

SPRINGER
REFERENCE

Azim Eskandarian
Editor

VOLUME 2

Handbook of Intelligent Vehicles

 Springer

Handbook of Intelligent Vehicles

Azim Eskandarian (Ed.)

Handbook of Intelligent Vehicles

With 627 Figures and 81 Tables



Editor

Azim Eskandarian
Center for Intelligent Systems Research
The George Washington University
801 22nd Street, NW, Phillips Hall 630
Washington, DC 20052
USA

ISBN 978-0-85729-084-7 e-ISBN 978-0-85729-085-4
DOI 10.1007/978-0-85729-085-4
ISBN Bundle 978-0-85729-086-1
Springer London Dordrecht Heidelberg New York

Library of Congress Control Number: 2011942207

© Springer-Verlag London Ltd. 2012

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use. The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my family for their love and understanding.
With special thanks to my sons, Saba and Kia for their dedicated assistance
and unceasing support.

Preface

Intelligent transportation systems (ITS) are defined as systems that use computers, controls, communications, and various automation technologies in order to enhance safety, throughput, and efficiency of transportation while reducing energy consumption and environmental impact. Although the scope of ITS is multimodal, road surface transportation has a major emphasis. Intelligent vehicles are obviously an integral part of ITS. Although the term “intelligent” is loosely defined in this context, it refers to incorporating a certain level of machine intelligence in the operation of a vehicle. The advancement of electronics, sensors, microprocessors, software, and embedded and electromechanical systems has enabled a significant level of automatic and autonomous functions in vehicles. Some of these functions are totally transparent to the driver and are triggered automatically, whereas others support the driver in the form of a driver assistance system.

The vehicle interacts with the driver, environment, and infrastructure. In intelligent vehicles, these interactions are augmented by the use of sensing, information exchange, and actuation of various primary or secondary driving tasks. These cover a broad range of functions from simple information exchanges to complex autonomous functions. The following are a few simple examples of existing systems or prototypes which improve the safety and efficiency of driving. A warning for an icy road or fog zone ahead enhances safety by providing timely information to the driver. This requires appropriate capacity to sense the environment by an ITS system and proper infrastructure to vehicle communications. In another example, a vehicle can sense the hazardous situation at hand and react automatically by enhancing the braking (e.g., in BAS) or traction and stability control to maintain its safe intended course within a lane. A radar- or vision-based collision avoidance system can sense obstacles ahead and prevent an imminent crash by automatically braking when the driver fails to do so. Vehicles’ energy consumption can be improved by increased knowledge of terrain and some intelligent shifting of transmission, or even through optimized trip planning using a smart navigation system. At another extreme, driverless vehicles can be driven autonomously to complete an entire trip from origin to destination while avoiding obstacles and obeying traffic laws.

Intelligent vehicles, as has been shown, cover a large and diverse range of technologies that span from dynamics of vehicles to information, communications, electronics, automation, human factors, etc. As such, research, development, and design of intelligent vehicles requires expertise and knowledge of various disciplines. Fortunately there are resources available within different scientific journals, conferences, and engineering professional societies that cover some aspects of intelligent vehicles, but they are very much field focused. For example, some journals cover control systems or vehicular dynamics. Other journals focus on communications or human factors, etc. Among engineering societies, IEEE (Institute for Electrical and Electronics Engineers) and

SAE (Society of Automotive Engineers) have specific divisions, conferences, and transactions that cover intelligent transportation systems and intelligent vehicles. There are also a few other journals dedicated to this topic. However, the scientific articles are typically focused on very specific problems and do not necessarily provide a broader picture or comprehensive coverage of large topics such as “intelligent vehicles.” There has been a lot of progress in the development of intelligent vehicles. Many systems are already on the market as high-end options in vehicles, and many other prototypes have been demonstrated in laboratory settings. Meanwhile, the development of intelligent vehicles is progressing rapidly. As consumers, intelligent vehicles, whether at their present state or at a future more autonomous state, affect our mobility and touch our everyday lives. Thus it is imperative that the cognizant scientific community provides the knowledge base and resources necessary for further developments.

Due to the diversity of technologies involved in intelligent vehicles, there aren’t any up-to-date books or references that provide the necessary coverage of this interesting topic. This handbook is intended to cover this gap. It should also be noted that due to the proprietary nature of developments in the industry, there is very little material in open literature with in-depth coverage of the science and methods underlying actual implemented technologies. An international team of editors and authors, each a recognized authority in his/her respective field of specialty, has been gathered to write about the most up-to-date topics concerning intelligent vehicles. Most authors of this handbook have conducted state-of-the-art research in each topic presented and hence provide the latest discoveries and methods.

The handbook is organized in an appropriate format to cover both the depth and breadth of this subject in 11 sections. Each section includes multiple chapters to cover each subject area. To the best of the editors’ knowledge, there are no other current resources with this depth and breadth available on the subject of intelligent vehicles.

This handbook serves vehicle engineers, scientists, researchers, students, and technical managers who are interested in the study, design, and development of intelligent vehicles. It is hoped that this handbook will serve the engineering, R&D, and academic community as a prominent resource for the foreseeable future. It is also hoped that it will be accepted by colleagues and students, helping them in their independent investigation of the topic.

Azim Eskandarian
Washington, DC, USA
January 2012

Acknowledgments

I would like to thank all Section Editors for their extensive efforts, dedication, flexibility, and valuable feedback in shaping and organizing this handbook. Each one of them is truly a distinguished expert in his/her field of specialty. In addition to their editorial services, many of them also authored or co-authored chapters, and thus added so much value to this Handbook.

My gratitude is especially extended to all authors for their contributions. They have selected and written some of the best material in their respective fields of specializations, unmatched anywhere in contemporary literature. The substance of their contributions has made this Handbook a truly unique resource for all readers.

I also would like to extend my sincere appreciations to all editorial and administrative staff at Springer for their hard work, steadfastness, and dedication to support authors and editors and for keeping production on schedule. Their diligent and unyielding efforts have made the timely production of this Handbook possible.

Last but not least, I would like to thank Springer-Verlag London Ltd. for approaching us to undertake this voluminous but worthy project.

With earnest hope that this Handbook will serve the R&D and engineering community well.

Azim Eskandarian
Editor-in-Chief

Biography



Dr. Azim Eskandarian, D.Sc.

Professor of Engineering and Applied Science
Director, Center for Intelligent Systems Research and
GW Transportation Safety and Security Area of
Excellence Program
School of Engineering and Applied Science
The George Washington University
Washington, DC, USA
eska@gwu.edu

Azim Eskandarian is a Professor of Engineering and Applied Science at The George Washington University (GWU). He is the founding director of the Center for Intelligent Systems Research (CISR) since 1996 and the director of the interdepartmental “Transportation Safety and Security” program, a major *Area of Excellence* at the university. Dr. Eskandarian co-founded the National Crash Analysis Center (NCAC) in 1992 and served as its Director from 1998 to 2002. He has almost three decades of R&D and engineering design experience in Dynamic Systems, Controls, Intelligent Systems, and Applied and Computational Mechanics, with applications in automotive engineering, transportation safety, intelligent vehicles, and robotics.

As Director of the CISR, his research had focused on intelligent vehicle systems for the future automotive and transportation safety and efficiency. He established four new research laboratories including a car- and a truck-driving simulator. His publications in the IEEE Transactions on Intelligent Transportation Systems are among the highest cited articles. He also served as General Chair of the 14th International IEEE Intelligent Transportation Systems Conference 2011.

Dr. Eskandarian’s pedagogical efforts have been instrumental in the establishment of a new and unique graduate program of study in “Transportation Safety” at GWU since 1994. Prior to joining GWU, Dr. Eskandarian was an assistant professor at Pennsylvania State University, York Campus. Before academia, he held engineering and project management positions in the defense industry (1983–1989).

Dr. Eskandarian is the author of over 135 articles, a book, and three edited volumes. He has been elected to the Board of Governors of IEEE ITS society twice since 2007 and was invited to the IEEE Committee on Transportation and Aerospace Policy in 2009. He is also active in the ASME Automotive and Transportation Systems Technical Committee of

Dynamic Systems and Control Division. He is the associate editor and editorial board member of five journals (including IEEE Transactions on ITS and IMech E. *Journal of Multi-body Dynamics*). In 2011, he was awarded the *Distinguished Researcher Award* of the GWU School of Engineering and Applied Science.

Dr. Eskandarian has served on several committees, boards, and review panels including DOT/NHTSA, NSF, NAC, TRB, and Canadian Centers of Excellence, and Canada Foundation for Innovation, and consulted for industry and government. He has been a member of ASME, IEEE, SAE, ITS America, and Sigma XI (2000–2003) professional society, and Tau Beta Pi and PI Tau Sigma Engineering Honor societies. He received his B.S. (with honors), M.S., and D.Sc. degrees in Mechanical Engineering from GWU, Virginia Polytechnic Institute and State University, and GWU, respectively.

List of Section Editors

Bart Van Arem

Delft University of Technology
Delft
The Netherlands

Azim Eskandarian

The George Washington University
Ashburn, VA
USA

Ernst Pucher

Institute for Powertrains and Automotive
Technology
Vienna University of Technology
Vienna
Austria

Alfred Pruckner

BMW Research and Development
Munich
Germany

Ralf Stroph

BMW Research and Development
Munich
Germany

Peter Pfeffer

Hochschule München
Munich
Germany

Peter Handel

Signal Processing Lab
ACCESS Linnaeus Center
Royal Institute of Technology
Sweden

Werner Huber

BMW Group
Munich
Germany

Alberto Broggi

Dipartimento di Ingegneria
dell'Informazione
Università di Parma
Parma
Italy

Klaus Kompass

BMW Group
Munich
Germany

Scott Andrews

Cogenia Partners, LLC
Petaluma, CA
USA

Christian Laugier

Renault
S.A.S.
Guyancourt
France

Michael Parent

Scientific Adviser
INRIA/IMARA
Rocquencourt
France

Table of Contents

List of Contributors xxi

Volume 1

1 Introduction to Intelligent Vehicles 1
Azim Eskandarian

Section 1 Overview of Intelligent Vehicle Systems and Approaches 15
Bart Van Arem

2 A Strategic Approach to Intelligent Functions in Vehicles 17
Bart van Arem

3 Sensing and Actuation in Intelligent Vehicles 31
Angelos Amditis · Panagiotis Lytrivis · Evangelia Portouli

4 Situational Awareness in Intelligent Vehicles 61
Zoltán Papp

5 Hierarchical, Intelligent and Automatic Controls 81
Bart De Schutter · Jeroen Ploeg · Lakshmi Dhevi Baskar · Gerrit Naus · Henk Nijmeijer

6 Behavioral Adaptation and Acceptance 117
Marieke H. Martens · Gunnar D. Jensen

7 Simulation Approaches to Intelligent Vehicles 139
Bart van Arem · Martijn van Noort · Bart Netten

Section 2 Vehicle Longitudinal and Lateral Control Systems 165
Azim Eskandarian

8 Vehicle Longitudinal Control 167
Jihua Huang

9 Adaptive and Cooperative Cruise Control 191
Fanping Bu · Ching-Yao Chan

10 Vehicle Lateral and Steering Control 209
Damoon Soudbakhsh · Azim Eskandarian

Section 3 Special Vehicular Systems 233
Ernst Pucher, Alfred Pruckner, Ralf Stroph and Peter Pfeffer

11 Drive-By-Wire 235
Alfred Pruckner · Ralf Stroph · Peter Pfeffer

12 Energy and Powertrain Systems in Intelligent Automobiles 283
Ernst Pucher · Luis Cachón · Wolfgang Hable

Section 4 Positioning, Navigation, and Trajectory Control 309
Peter Handel

13 Global Navigation Satellite Systems: An Enabler for In-Vehicle Navigation 311
John-Olof Nilsson · Dave Zachariah · Isaac Skog

14 Enhancing Vehicle Positioning Data Through Map-Matching 343
Mohammed A. Quddus · Nagendra R. Velaga

15 Situational Awareness and Road Prediction for Trajectory Control Applications 365
Christian Lundquist · Thomas B. Schön · Fredrik Gustafsson

16 Navigation and Tracking of Road-Bound Vehicles Using Map Support 397
Fredrik Gustafsson · Umut Orguner · Thomas B. Schön · Per Skoglar · Rickard Karlsson

17 State-of-the-Art In-Car Navigation: An Overview 435
Isaac Skog · Peter Händel

18 Evolution of in-car Navigation Systems 463
Koichi Nagaki

Section 5 Driver Assistance	489
<i>Azim Eskandarian</i>	
19 Fundamentals of Driver Assistance	491
<i>Azim Eskandarian</i>	
20 Driver Behavior Modeling	537
<i>Samer Hamdar</i>	
21 Using Naturalistic Driving Research to Design, Test and Evaluate Driver Assistance Systems	559
<i>Gregory M. Fitch · Richard J. Hanowski</i>	
22 Intelligent Speed Adaptation (ISA)	581
<i>Jeremy J. Blum · Azim Eskandarian · Stephen A. Arhin</i>	
Section 6 Safety and Comfort Systems	603
<i>Werner Huber and Klaus Kompass</i>	
23 Safety and Comfort Systems: Introduction and Overview	605
<i>Klaus Kompass · Werner Huber · Thomas Helmer</i>	
24 Adaptive Cruise Control	613
<i>Hermann Winner</i>	
25 Forward Collision Warning and Avoidance	657
<i>Markus Maurer</i>	
26 Lane Departure and Lane Keeping	689
<i>Jens E. Gayko</i>	
27 Integral Safety	709
<i>Klaus Kompass · Christian Domsch · Ronald E. Kates</i>	
28 Lane Change Assistance	729
<i>Arne Bartels · Marc-Michael Meinecke · Simon Steinmeyer</i>	
29 Steering and Evasion Assist	759
<i>Thao Dang · Jens Desens · Uwe Franke · Dariu Gavrilă · Lorenz Schäfers · Walter Ziegler</i>	

Volume 2

30 Proactive Pedestrian Protection 785
Stefan Schramm · Franz Roth · Johann Stoll · Ulrich Widmann

31 Parking Assist 829
Michael Seiter · Hans-Jörg Mathony · Peter Knoll

32 Post-crash Support Systems 865
Jeffrey S. Augenstein · George T. Bahouth

33 Map Data for ADAS 881
John Craig

Section 7 Drowsy and Fatigued Driver Detection,
Monitoring, Warning 893
Azim Eskandarian

34 Advances in Drowsy Driver Assistance Systems Through
Data Fusion 895
Darrell S. Bowman · William A. Schaudt · Richard J. Hanowski

35 Drowsy Driver Posture, Facial, and Eye Monitoring Methods 913
Jixu Chen · Qiang Ji

36 Drowsy and Fatigued Driving Problem Significance and Detection
Based on Driver Control Functions 941
Azim Eskandarian · Ali Mortazavi · Riaz Akbar Sayed

37 Drowsy and Fatigued Driver Warning, Counter Measures,
and Assistance 975
Riaz Akbar Sayed · Azim Eskandarian · Ali Mortazavi

Section 8 Vision-based Systems 997
Alberto Broggi

38 Image Processing for Vehicular Applications 999
Massimo Bertozzi

39 Camera Technologies 1011
Paolo Grisleri

40	Perception Tasks: Lane Detection	1021
	<i>Luca Mazzei · Paolo Zani</i>	
41	Perception Tasks: Obstacle Detection	1033
	<i>Stefano Debattisti</i>	
42	Perception Tasks: Traffic Sign Recognition	1043
	<i>Pier Paolo Porta</i>	
43	Vision-Based ACC	1061
	<i>Matteo Panciroli</i>	
44	Vision-Based Blind Spot Monitoring	1071
	<i>Elena Cardarelli</i>	
Section 9 Vehicular Communications Systems		1089
	<i>Scott Andrews</i>	
45	Vehicular Communications Requirements and Challenges	1091
	<i>Scott Andrews</i>	
46	Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) Communications and Cooperative Driving	1121
	<i>Scott Andrews</i>	
47	Probes and Intelligent Vehicles	1145
	<i>Christopher Wilson</i>	
48	Threat Model, Authentication, and Key Management	1173
	<i>Stan Pietrowicz</i>	
49	Security, Privacy, Identifications	1217
	<i>William Whyte</i>	
Section 10 Fully Autonomous Driving		1269
	<i>Christian Laugier</i>	
50	Autonomous Driving: Context and State-of-the-Art	1271
	<i>Javier Ibañez-Guzmán · Christian Laugier · John-David Yoder · Sebastian Thrun</i>	
51	Modeling and Learning Behaviors	1311
	<i>Dizan Vasquez · Christian Laugier</i>	

52 Vision and IMU Data Fusion: Closed-Form Determination of the Absolute Scale, Speed, and Attitude 1335
Agostino Martinelli · Roland Siegwart

53 Vision-Based Topological Navigation: An Implicit Solution to Loop Closure 1355
Youcef Mezouar · Jonathan Courbon · Philippe Martinet

54 Awareness of Road Scene Participants for Autonomous Driving 1383
Anna Petrovskaya · Mathias Perrollaz · Luciano Oliveira · Luciano Spinello · Rudolph Triebel · Alexandros Makris · John D. Yoder · Christian Laugier · Urbano Nunes · Pierre Bessiere

55 Iterative Motion Planning and Safety Issue 1433
Thierry Fraichard · Thomas M. Howard

56 Risk Based Navigation Decisions 1459
Anne Spalanzani · Jorge Rios-Martinez · Christian Laugier · Sukhan Lee

57 Probabilistic Vehicle Motion Modeling and Risk Estimation 1479
Christopher Tay · Kamel Mekhnacha · Christian Laugier

Section 11 A Look to the Future of Intelligent Vehicles 1517
Michael Parent

58 Legal Issues of Driver Assistance Systems and Autonomous Driving 1519
Tom Michael Gasser

59 Intelligent Vehicle Potential and Benefits 1537
Claude Lorangeau

60 Applications and Market Outlook 1553
Michel Parent

Index 1575

List of Contributors

Angelos Amditis

Institute of Communication and
Computer Systems (ICCS)
Zografou, Athens
Greece

Scott Andrews

Systems Engineering and Intellectual
Property Management in the
Automotive, Mobile Computing, and
Communications Domains
Cogenia Partners, LLC
Petaluma, CA
USA

Bart van Arem

Civil Engineering and Geoscience
Transport & Planning
Delft University of Technology
Delft, CN
The Netherlands

Stephen A. Arhin

Civil Engineering
Howard University
NW, Washington, DC
USA

Jeffrey S. Augenstein

Miller School of Medicine
University of Miami
Florida
USA

George T. Bahouth

Impact Research, Inc.
Columbia
USA

Arne Bartels

Driver Assistance and Integrated Safety
Volkswagen Group Research
Wolfsburg
Germany

Massimo Bertozzi

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Pierre Bessiere

INRIA and Collège de France
Saint Ismier Cedex
France

Jeremy J. Blum

Computer Science
Penn State University Harrisburg
Middletown, PA
USA

Darrell S. Bowman

Center for Truck and Bus Safety
Virginia Tech Transportation Research
Institute
Transportation Research Plaza
Blacksburg, VA
USA

Fanping Bu

California PATH, Institute of
Transportation Studies
University of California at Berkeley Bldg.
Richmond Field Station
Richmond, CA
USA

Luis Cachón

Institute for Powertrains and Automotive
Technology
Vienna University of Technology
Vienna
Austria

Elena Cardarelli

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Ching-Yao Chan

California PATH, Institute of
Transportation Studies
University of California at Berkeley Bldg.
Richmond Field Station
Richmond, CA
USA

Jixu Chen

Visualization and Computer Vision Lab
GE Global Research Center
Niskayuna, NY
USA

Jonathan Courbon

Clermont Université
Université Blaise Pascal
LASMEA
and
CNRS
LASMEA

John Craig

JCC
Munich
Germany

Thao Dang

Group Research and Advanced
Engineering
Driver Assistance and Chassis Systems
Daimler AG
Sindelfingen
Germany

Bart De Schutter

Delft Center for Systems and Control
Delft University of Technology
Delft
The Netherlands

Stefano Debattisti

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Jens Desens

Group Research and Advanced
Engineering
Driver Assistance and Chassis Systems
Daimler AG
Sindelfingen
Germany

Lakshmi Dhevi Baskar

Delft Center for Systems and Control
Delft University of Technology
Delft
The Netherlands

Christian Domsch

BMW Group
Munich
Germany

Azim Eskandarian

Center for Intelligent Systems Research
The George Washington University
Washington, DC
USA

Gregory M. Fitch

Virginia Tech Transportation Institute-
Truck and Bus Safety
Blacksburg, VA
USA

Thierry Fraichard

INRIA Grenoble - Rhône-Alpes
CNRS-LIG and Grenoble University
Grenoble
France

Uwe Franke

Group Research and Advanced
Engineering
Driver Assistance and Chassis Systems
Daimler AG
Sindelfingen
Germany

Tom Michael Gasser

F4 -Co-operative Traffic and Driver
Assistance Systems
Federal Highway Research Institute
(BASt)
Bergisch Gladbach
Germany

Dariu Gavrila

Group Research and Advanced
Engineering
Driver Assistance and Chassis Systems
Daimler AG
Ulm
Germany

Jens E. Gayko

VDE Association for Electrical, Electronic
and Information Technologies
VDE Headquarters
Frankfurt am Main
Germany

Paolo Grisleri

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Fredrik Gustafsson

Division of Automatic Control
Department of Electrical Engineering
Linköping University
Linköping, SE
Sweden

Wolfgang Hable

Institute for Powertrains and Automotive
Technology
Vienna University of Technology
Vienna
Austria

Samer Hamdar

The George Washington University
Ashburn, VA
USA

Peter Händel

School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm
Sweden

Richard J. Hanowski

Center for Truck and Bus Safety
Virginia Tech Transportation Research
Institute
Transportation Research Plaza
Blacksburg, VA
USA

Thomas Helmer

BMW Group
Munich
Germany

Thomas M. Howard

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA
USA

Jihua Huang

Partners for Advanced Transportation
Technologies (PATH)
Institute of Transportation Studies
University of California at Berkeley
Richmond, CA
USA

Werner Huber

BMW Group
Munich
Germany

Javier Ibañez-Guzmán

Multimedia and Driving Assistance
Systems
Renault S.A.S
Guyancourt
France

Gunnar D. Jenssen

Transport Research
SINTEF
Trondheim
Norway

Qiang Ji

Department of Electrical, Computer, and
Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY
USA

Rickard Karlsson

Division of Automatic Control
Department of Electrical Engineering
Linköping University
Linköping, SE
Sweden

Ronald E. Kates

REK-Consulting
Otterfing
Germany

Peter Knoll

Karlsruhe Institute of Technology (KIT)
Karlsruhe
Germany

Klaus Kompass

BMW Group
Munich
Germany

Christian Laugier

e-Motion Project-Team
INRIA Grenoble Rhône-Alpes
Saint Ismier Cedex
France

Claude Lurgeau

Mines ParisTech
Paris
France

Sukhan Lee

School of Information and
Communication Engineering
Department of Interaction Science
ISRI (Intelligent Systems
Research Institute)
Sungkyunkwan University
Jangan-gu, Suwon
Rep. of Korea (South Korea)

Christian Lundquist

Department of Electrical Engineering
Division of Automatic Control
Linköping University
Linköping, SE
Sweden

Panagiotis Lytrivis

Institute of Communication and
Computer Systems (ICCS)
Zografou, Athens
Greece

Alexandros Makris

INRIA
Saint Ismier Cedex
France

Marieke H. Martens

TNO
Soesterberg
The Netherlands
and
University of Twente
Enschede
The Netherlands

Agostino Martinelli

INRIA
INRIA Rhone Alpes
avenue de l'Europe
Grenoble, Montbonnot, Saint Ismier
Cedex
France

Philippe Martinet

Clermont Université
IFMA
LASMEA
and
CNRS
LASMEA

Hans-Jörg Mathony

Robert Bosch GmbH
Leonberg
Germany

Markus Maurer

Technische Universität Braunschweig
Institut für Regelungstechnik
Braunschweig
Germany

Luca Mazzei

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Marc-Michael Meinecke

Driver Assistance and Integrated Safety
Volkswagen Group Research
Wolfsburg
Germany

Kamel Mekhnacha

Probayes SAS
Inovalée Saint Ismier Cedex
France

Youcef Mezouar

Clermont Université
Université Blaise Pascal
LASMEA
and
CNRS
LASMEA

Ali Mortazavi

Partners for Advanced Transportation
Technology (PATH)
University of California
Berkeley, CA
USA

Koichi Nagaki

Software R&D Department
Car Electronics Engineering Division
Pioneer Corporation
Kawagoe-Shi, Saitama-Ken
Japan

Gerrit Naus

Department of Mechanical Engineering
Dynamics and Control Section
Eindhoven University of Technology
Eindhoven
The Netherlands

Bart Netten

Netherlands Organization for Applied
Scientific Research TNO
Delft
The Netherlands

Henk Nijmeijer

Department of Mechanical Engineering
Dynamics and Control Section
Eindhoven University of Technology
Eindhoven
The Netherlands

John-Olof Nilsson

School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm
Sweden

Martijn van Noort

Netherlands Organization for Applied
Scientific Research TNO
Delft
The Netherlands

Urbano Nunes

Faculty of Science and Technology
Coimbra University
Pólo II
Coimbra
Portugal

Luciano Oliveira

Faculty of Science and Technology
Coimbra University
Pólo II
Coimbra
Portugal

Umut Orguner

Division of Automatic Control,
Department of Electrical Engineering
Linköping University
Linköping, SE
Sweden

Matteo Pancioli

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Zoltán Papp

TNO Technical Sciences
The Hague
Netherlands

Michel Parent

Unité de recherche INRIA Paris
Rocquencourt
Project IMARA
Le Chesnay Cedex
France

Mathias Perrollaz

e-Motion Project-Team
INRIA Grenoble Rhône-Alpes
Saint Ismier Cedex
France

Anna Petrovskaya

Artificial Intelligence Laboratory
Stanford University
Stanford, CA
USA

Peter Pfeffer

Hochschule München
Munich
Germany

Stan Pietrowicz

Telcordia Technologies
Red Bank, NJ
USA

Jeroen Ploeg

Automotive
TNO Technical Sciences
Helmond
The Netherlands

Pier Paolo Porta

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Evangelia Portouli

National Technical University of Athens
Zografou, Athens
Greece

Alfred Pruckner

BMW Research and Development
Munich
Germany

Ernst Pucher

Institute for Powertrains and Automotive
Technology
Vienna University of Technology
Vienna
Austria

Mohammed A. Quddus

Department of Civil and Building
Engineering
Loughborough University
Leicestershire
UK

Jorge Rios-Martinez

e-Motion Project-Team
INRIA Rhône-Alpes
Saint Ismier Cedex
France

Franz Roth

Vehicle Safety Development
AUDI AG
Ingolstadt
Germany

Riaz Akbar Sayed

Mechanical Department
NWFP University of Engineering and
Technology
Peshawar, North West Frontier Provi
Pakistan

William A. Schaudt

Center for Truck and Bus Safety
Virginia Tech Transportation Research
Institute
Transportation Research Plaza
Blacksburg, VA
USA

Lorenz Schäfers

Group Research and Advanced
Engineering
Driver Assistance and Chassis Systems
Daimler AG
Sindelfingen
Germany

Thomas B. Schön

Division of Automatic Control
Department of Electrical Engineering
Linköping University
Linköping, SE
Sweden

Stefan Schramm

Vehicle Safety Development
AUDI AG
Ingolstadt
Germany

Michael Seiter

Robert Bosch GmbH
Leonberg
Germany

Roland Siegwart

Inst.f. Robotik u. Intelligente Systeme
ETHZ, Zurich
Zurich
Switzerland

Isaac Skog

School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm
Sweden

Per Skoglar

Division of Automatic Control
Department of Electrical Engineering
Linköping University
Linköping, SE
Sweden

Damoon Soudbakhsh

Center for Intelligent Systems Research
The George Washington University
Washington, DC
USA

Anne Spalanzani

UPMF-Grenoble 2/INRIA Rhône-Alpes/
Lig UMR
Grenoble
France

Luciano Spinello

University of Freiburg
Freiburg
Germany
and
Inst.f. Robotik u. Intelligente Systeme
ETH Zurich
Zurich
Switzerland

Simon Steinmeyer

Driver Assistance and Integrated Safety
Volkswagen Group Research
Wolfsburg
Germany

Johann Stoll

Vehicle Safety Development
AUDI AG
Ingolstadt
Germany

Ralf Stroph

BMW Research and Development
Munich
Germany

Christopher Tay

Probayes SAS
Inovalée Saint Ismier Cedex
France

Sebastian Thrun

Stanford University
Stanford, CA
USA

Rudolph Triebel

University of Oxford
Oxford
UK
and
Inst.f. Robotik u. Intelligente Systeme
ETH Zurich
Zurich
Switzerland

Dizan Vasquez

ITESM Cuernavaca
Mexico

Nagendra R. Velaga

ITS, dot.rural Digital Economy Research
Hub
University of Aberdeen
UK

William Whyte

Security Innovation
Wilmington, MA
USA

Ulrich Widmann

Vehicle Safety Development
AUDI AG
Ingolstadt
Germany

Christopher Wilson

TomTom Group
San Francisco Bay Area, CA
USA

Hermann Winner

Technische Universität Darmstadt
Fachgebiet Fahrzeugtechnik
Darmstadt
Germany

John-David Yoder

Mechanical Engineering Department
Ohio Northern University
Ada, OH
USA

Dave Zachariah

School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm
Sweden

Paolo Zani

Dip. Ing. Informazione
Università di Parma
Parma
Italy

Walter Ziegler

Group Research and Advanced
Engineering, Driver Assistance and
Chassis Systems
Daimler AG
Sindelfingen
Germany

1 Introduction to Intelligent Vehicles

Azim Eskandarian
Center for Intelligent Systems Research, The George Washington
University, Washington, DC, USA

- 1 Background 2
- 2 Traffic Safety 2
- 3 Energy and Environment 4
- 4 Governmental Efforts: Intelligent Transportation Systems 5
- 5 Some Basic Definitions of Intelligent Vehicles 7
- 6 Purpose and Scope of This Handbook 9
 - 6.1 Purpose 9
 - 6.2 Coverage of Topics 10

1 Background

Human mobility was revolutionized by the advent of automobiles. As the number of automobiles increased, streets and roads were expanded to accommodate inter- and intracity travel and rural connectivity. Concurrently, traffic regulations were developed to control orderly routing of vehicles and ensure safety. With the advancement of technology, vehicle power, performance, and travel range have improved tremendously to a level that they have even reshaped the city sprawl and the way people live today. Throughout the years, vehicles have evolved to become sophisticated technological machines that extend mobility to leisure, comfort, luxury, sports, and, for some, an extension and expression of their image and personality. As with any other technology, this advancement in mobility has brought its own challenges in safety, pollution, and energy demands. As vehicles have become a necessity of life, encompassing almost every aspect of our daily lives, throughout the years various laws, regulations, and standards have been developed for vehicle design, production, and interface with the environment and the driver. Today, an enormous amount of engineering design, prototyping, testing, evaluation, and redesign goes into the production of a vehicle to ensure that the intended vehicular performance, safety, energy, and environmental requirements are met. In addition to all of these requirements, a significant aspect of vehicular design concerns the interaction of driver with vehicle. This interaction includes the primary tasks of controlling the vehicle and all the auxiliary and secondary tasks normally performed by the driver.

Although the driver has remained largely in control of driving vehicles since the inception of the automobile, several systems have been introduced to enhance vehicle response in braking and handling. Advances in microprocessors and computers have had a tremendous impact on vehicle design, but their full potential is yet to be realized. Today, vehicles have many sensors and electronic systems that contribute to automatic control of subsystems for a range of functions from controlling vehicle dynamics (ABS and traction control) to supporting the driver in trip planning and route selection (e.g., navigation systems). Intelligent vehicles aim to fully utilize available technologies to assist drivers by enhancing handling, safety, efficiency, and the comfort of driving. What constitutes “intelligence” in the automotive context is further described in the following sections.

This chapter provides a brief and introductory overview of intelligent vehicles. It highlights the significant safety challenges of traffic crashes and introduces the enormous energy demands of the vehicular transportation sector in the USA. Then, samples of governmental efforts in developing intelligent vehicles within the scope of intelligent transportation systems in the USA and abroad are described. Some basic definitions of what constitutes an intelligent vehicle are provided. Finally the purpose, scope, and coverage of the topics in this handbook are explained.

2 Traffic Safety

Vehicle safety remains a major human challenge. In 2009, traffic crashes caused 33,808 fatalities and 2.2 million injuries in the USA, resulting in an estimated economic loss of

\$230.6 billion (Traffic Safety Facts 2009, 2010). Traffic injuries remain the leading cause of death for teenagers in the USA. The rest of the world suffers from similar alarming statistics. The World Health Organization (WHO) reports 1.2 million fatalities in traffic accidents and 20–50 million injuries each year worldwide (Global status report on road safety; WHO 2009). One half of the fatalities involve what the WHO refers to as “vulnerable road users”: pedestrians, cyclists, or motorized two-wheelers. In the USA, 35% of reported deaths are among “vulnerable road users.”

While direct statistics reveal the enormity of the problem, the fatality rate, which is the number of fatalities on a per-vehicle-mile-driven basis, constitutes a more accurate means of measuring highway safety trends. The number of fatalities per 100 million-vehicle-miles-traveled in the USA was 1.13 during 2009 and the number of fatalities per 100,000 population was 11.01. Since the actual vehicle miles traveled each year are largely based on crude estimates, even these rates are not the most suitable for detailed analysis of safety, providing instead a general metric on trends over many years.

Worldwide, road traffic injuries rank ninth, at 2.2%, among leading causes of death and it is estimated to rise to rank 5 by 2030, according to a 2009 World Health Organization report (Global status report on road safety; WHO 2009). The same report shows road traffic injuries as the first, second, and third leading causes of death among all diseases for age ranges 15–29, 5–14, and 30–44 years, respectively. This makes teenagers one of the most vulnerable populations. According to the WHO, addressing this epidemic safety problem in a comprehensive manner requires strategies, involvement, and coordinated responses between multiple governmental agencies (transport sector, police, health, etc.). It also necessitates finances for planned activities with specific goals and targets to reduce fatalities and injuries over a given period. However, only one third of all countries have such a national road safety strategy that is endorsed by the government. Although the highest rates of fatality and injury are seen in low- and middle-income countries, hazard of driving remains a persisting problem in high-income countries as well. As a sample comparison, the death rate per 100,000 people in the USA in 2004 was 13.9, while it was 6, 6.7, and 5.2 for Germany, China, and Japan, respectively. Even in countries with better records of road safety (such as the Netherlands, Sweden, or the United Kingdom for instance), there remains considerable room for improvement. For example, in Sweden (which has one of the best safety records in the world), the road transport system is notably responsible for causing more child fatalities than other forms of transportation such as rail and air travel; traffic crashes alone in Sweden are responsible for 20% of deaths in children aged 5–19 (Global status report on road safety; WHO 2009).

It should also be noted that fatality and injury rates have been declining steadily for many years. According to NHTSA (Traffic Safety Facts Research Note 2010), passenger car occupant fatalities declined for the seventh consecutive year in 2009, and are now at their lowest level since NHTSA began collecting fatality crash data in 1975. Light-truck occupant fatalities have dropped for the fourth consecutive year, and are at their lowest level since 1997. On the other hand, motorcycle fatalities have been increasing for 11 consecutive years, with the exception of 2009, when it declined by 24%, then accounting for 13% of traffic fatalities. Fatalities per 100 million-vehicle-miles-traveled

in the USA have decreased almost steadily from 1.73 in 1997 to 1.13 in 2009. Similar declining trends are observed in other high-income countries. These could perhaps be attributed to enhanced safety technologies in vehicles, better roads, persuasive safety campaigns, and vigorous law enforcement resulting in patterns such as increased belt use, reduced speed, reduced frequency of driving under influence, etc.

Despite these positive trends, the statistics are still alarming and require much more attention to reach a goal of near zero or zero fatality in road transport. The WHO also stresses that although death rates have been declining over the past five decades in high-income countries, road traffic injuries still remain an important cause of death and disability. A system approach that considers human-vehicle-environment interactions with contributing countermeasures in precrash, crash, and postcrash temporal modes is required to tackle this difficult and challenging problem. Such an approach was envisioned by the first NHTSA Administrator, William Haddon. This strategy is based on the premise that humans make mistakes and that driver error and impairment is responsible for the majority of crashes. Therefore, in addition to government programs, public education, and enforcement campaigns, technological solutions have been developed to remedy the problem. The automotive industry has made significant strides in this respect, and governments around the world have increased safety regulations.

Vehicle safety is largely driven by the two forces of consumer demand and government regulation. Passive and active systems have been developed throughout the years to enhance vehicle safety. With the invention and incorporation of seatbelts in the 1960s, crush zone in the 1970s, airbags in the 1980s, and smart airbags in the 2000s, passive safety systems have largely improved the crashworthiness of vehicles. In a similar, parallel trend, design of active safety systems such as ABS in the 1970s, traction control in the 1980s, electronic stability control and brake assist in the 1990s, adaptive cruise control, blind spot detection, and lane departure detection in the 2000s have contributed significantly to safety improvements in road transport. More recent technologies like pedestrian detection and concepts like integrated safety, along with the forthcoming communications, driver assistance, and autonomous (driverless) driving in intelligent vehicles are expected to bring the next wave of improvements in vehicular safety over many years to come.

3 Energy and Environment

Another major challenge facing the transport system is that of energy requirements and environmental impacts. With the exception of a few years of economic downturn, there has been a steady, long-term, global increase in both the number of vehicles and the amount of miles traveled each year. Energy for the transport industry is largely supplied by fossil fuels, gasoline, diesel, and gas, with the latter having a smaller portion of the market. The following data are taken from various tables of National Transportation Statistics ([National Transportation Statistics web site](#)). In 2008, the USA accounted for 22.7% of the world's petroleum use (and about 25% from 1990 to 2005 and 24% from 2005 to 2007). In 2008, 70.4% of the total domestic petroleum consumption in the USA was used for

transportation (all modes). Similarly in 2009 and 2010, the transportation sector used 71.6% and 71.3%, respectively, of total petroleum in the USA.

In 2009, the US transportation sector (including all modes) consumed 28.4% (26.94 Quadrillion Btu) of the total energy consumption of the country (94.72 Quadrillion Btu). Ninety-four percent of primary demand in the transportation sector was met by petroleum ([National Transportation Statistics Web site](#), Tables 4–3). In 2009, 254.213 million registered motor vehicles consumed 168,140 million gallons of fuel in the USA, resulting in 661 gallons of average fuel consumed by each vehicle. On average, US transportation fuel consumption accounts for over 70% of total US oil consumption, and more than 65% of that amount is for personal vehicles, that is, personal vehicles consume about 45.5% of total US oil consumption ([Energy Independence web site](#)).

With the continued economic growth of large population countries like China, India, etc., a much greater rise in energy demand is anticipated worldwide, particularly for the transport sector. The ever-increasing demand for energy will inevitably pose a major concern for rising fuel costs in the not too distant future. Because of this, most developed countries have been looking into alternative resources for their energy needs. A considerable trend toward either hybrid or totally electric vehicles has been observed in recent years among other renewable fuel alternatives. Despite the significant progress made to date in electric vehicle technology, it still does not provide the range and power of gasoline or diesel-fueled vehicles. However, much research is in progress in this field. The necessary infrastructure and smart electric grids also need to be in place to address the charging demands of electric vehicles. Without suitably designed infrastructure resources for charging in the long term, vehicle electrification may remain a limited success restricted only to compact urban areas. Electrification also brings the significant environmental benefit of emissions reduction.

Intelligent vehicles can contribute to reduction of energy consumptions and environmental impacts through various methods, for example, by servo-level controls of power components and regenerative drive systems (by continuous monitoring of dynamics of the vehicle and road), as well as by better energy conscious higher level decisions in route and trip selections.

4 Governmental Efforts: Intelligent Transportation Systems

Intelligent vehicles, as part of the larger transportation system, are aimed to address these challenging problems of safety and fuel economy, among other goals. In addition to the private-sector and automotive industry, governments in all developed countries have launched programs in their transportation departments (ministries) to bring automation and advanced technologies of information, computers, control, command, and communications to various elements of their transportation infrastructure since the 1980s. The goals of these programs are to reduce congestion, improve safety and comfort of travel, increase energy efficiency, and reduce environmental pollution caused by the transport sector.

In the US DoT, these efforts are conducted under a national intelligent transportation systems (ITS) program, earlier known as IVHS (intelligent vehicle highway systems), which was introduced during the 1980s, resulted in projects during the 1990s, and has continued ever since. In the mid-1980s, a group of Federal and State Governments, universities, and private-sector companies, which took the name Mobility 2000 by 1988, started discussions of introducing advanced technologies in future transportation systems (Proceedings of a National Workshop on IVHS Sponsored by Mobility 2000 [1990](#)). This later resulted in the formation of IVHS America, a professional advocacy organization, which was renamed to ITS America, as the scope of IVHS was expanded to include intermodal transportation. Almost concurrently with the USA, The Commission of European Communities (CEC) started programs in *road transport informatics* (RTI) and *integrated road transport environment* (IRTE). Starting in 1989, two major research and development projects entitled DRIVE I and DRIVE II were carried out under the modern European designation of Advanced Transport Telematics (ATT) (Catling [1993](#)). Japan and many other countries from East Asia and South Pacific have started similar programs to battle the rising traffic deaths, injuries, and congestion problems. Since the 1990s, a large number of R&D and operational test and evaluation projects have been completed and numerous intelligent transportation systems have been implemented worldwide.

In the USA, the ITS program was further shaped into R&D and operational test projects through the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA) with a \$1.2 billion authorization for 5 years, 1992–1997. In addition to R&D and evaluation, ISTEA established a federal program to also promote the implementation of Intelligent Transportation Systems (ITS). Subsequently, the ITS program continued with a similar level of funding through two additional bills passed by congress, namely, the Transportation Efficiency Act for the 21st Century (TEA-21) through 2003, and the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). The ITS Deployment Program ended at the close of fiscal year 2005, but SAFETEA-LU legislation continued ITS research at \$110 million annually through fiscal year 2009. In addition to authorized ITS funding, ITS projects are eligible for regular federal-aid highway funding ([ITS, DoT JPO Web site](#)). Currently, efforts are underway by US DoT under its 5-year ITS Strategic Research Plan 2010–2014, which focuses heavily on the wireless connectivity of vehicles and vehicle-to-infrastructure communications systems to provide hazard information and alerts to drivers in order to help prevent crashes and enhance safety and mobility while reducing environmental impacts.

Since its inception, the US ITS program has identified a set of user services and functions that required to be developed mostly through public and private partnerships by the federal, state, and local transportation agencies, toll and transit authorities, and their private and commercial counterparts. For example, advanced transportation management systems (ATMS), advanced traveler information systems (ATIS), advanced vehicle control and safety systems (AVCS), advanced public transportation systems (APTS), commercial vehicle operations (CVO), automatic toll collection, and emergency response systems are among the user services defined within the US ITS program. A national *ITS Architecture* has also been developed to better define the functionalities

and interrelationship of various ITS systems and components, and accommodate the integration of future systems. ITS training has also been an integral part of these efforts. Almost, all US DoT administrations including FHWA, NHTSA, FMCSA, FTA, RITA, etc., have ITS programs, coordinated through an ITS joint program office (JPO) under US DOT's research and innovation technology administration (RITA).

Intelligent vehicles are an integral element of many ITS systems. What is known today as intelligent vehicles is an outcome of some of the aforementioned programs and the market-driven efforts of the automotive industry to improve vehicle safety and efficiency. Within national programs, many specific ITS projects focusing on vehicles, driver-vehicle interface, and vehicle-infrastructure integration have contributed to the development of intelligent vehicles. The DoT and industry-sponsored automated highway systems (AHS) in 1990s, intelligent vehicle initiative (IVI) and vehicle-infrastructure integration (VII) during late 1990s and early 2000s, *Intellidrive* program in the mid-2000s, and more recently ITS *connected vehicle research* have been major initiatives, among many more DoT agency-specific projects, that helped shape the progress in intelligent vehicles. For example, they resulted in a successful demonstration of vehicle platoon formation in 1997 (AHS program), development, testing, and evaluation of collision warning technologies, and vehicle-to-vehicle and vehicle-to-infrastructure communications for safety alerts and intersection collision avoidance, among others.

5 Some Basic Definitions of Intelligent Vehicles

The term “intelligent” is defined by Webster's Dictionary as “having or indicating a high or satisfactory degree of intelligence and mental capacity,” and “intelligence” is “the ability to learn or understand or to deal with new or trying situations: reason” ([Web Site: Merriam Webster, On-line](#)). Carrying this definition to road vehicles, however, raises the expectations for such vehicles too high. Designing machines with intelligence has been the main goal of the fields of artificial intelligence and robotics for several decades. Although significant progress has been made in these fields, regarding demonstration of true human-like intelligence, only limited success is achieved in certain tasks and manipulations under well-defined conditions. The term “intelligent” in the context of machines is defined by Webster's Dictionary as “guided or controlled by a computer; *especially*: using a built-in microprocessor for automatic operation, for processing of data, or for achieving greater versatility.” This definition seems more appropriate for intelligent vehicles, which are viewed as machines controlled partly by humans and partly by computers (microprocessor). The focus of this handbook is on intelligent land vehicles and specifically on road vehicles, that is, passenger cars, buses, trucks, etc. In this context, a more general and technically pragmatic definition of *intelligent vehicles* is a vehicle that performs certain aspects of driving either autonomously or assists the driver to perform his/her driving functions more effectively, all resulting in enhanced safety, efficiency, and environmental impact.

The term “autonomous” implies that a machine has the intelligence to carry out a task without human (operator) guidance. Tasks could be manipulating a machine, executing

a command, performing a weld or placing parts (e.g., in case of industrial robots), piloting a plane, or driving a car. Performing tasks autonomously, therefore, requires that a machine has a desired goal to achieve, can sense or detect the situation at hand, perceive the conditions, analyze and make decisions for actions, and execute an action or response. These are the cognitive and motor actions (also known as perception-response) that humans do continuously during driving a vehicle while interacting with the roadway, environment, and traffic, and obeying the traffic signals and regulations. In case of driving, these steps translate to, for example, a desired goal to maintain within certain speed and lane boundaries, detect the surrounding traffic, perceive the situation if obstacles or other vehicles emerge on the path, analyze and make a decision to react, and finally execute the tasks of braking, steering, or accelerating as necessary. An intelligent vehicle performs these maneuvering functions autonomously or helps the driver to better control the vehicle. In a sense, the intelligent vehicle substitutes or assists the role of the human in the human-vehicle-environment interactions during driving.

In addition to the trajectory control and navigation described above, the general tasks of driving include many other functions, such as trip planning, route selection, responses to traffic and alternate route choices, economic consideration, vehicle condition monitoring, etc. In the broad sense of definition, an intelligent vehicle provides support or assistance in all of these functions as much as possible.

The term “ADAS” (advanced driver assistance systems) refers primarily to the vehicle handling functions that an intelligent vehicle provides either autonomously or supports the driver to execute more effectively. The term “driver assistance”, although sometimes used interchangeably with ADAS, refers not only to the vehicle handling but also to all other in-vehicle support systems for the driver, namely, all functions that assist a driver to execute a trip (including planning and reaching from origin to destination) safer, more efficient, and with less harmful environmental impacts.

Intelligent vehicles also improve the energy efficiency by reducing power consumption. This is done by introducing the necessary *intelligence* in the power train. The reduction in consumption is achieved through a multitude of approaches. At the higher trip planning level, an optimal trip choice minimizes distance or time, hence reducing the energy consumption. At trajectory level, the elevation, drag, and other road and environment conditions can be sensed, and the acceleration and velocity of the vehicle controlled to optimize fuel consumption. Close sensing and monitoring of the engine, power train, and transmission components aids in optimal performances that reduce consumptions. Some of these capabilities, packaged as *eco driving* mode, are now commercially available.

In summary, intelligent vehicles support driving functions at various levels of automation. At the highest level of automation, an intelligent vehicle can drive autonomously (driverless), that is, without intervention of the driver in all aspects of trip planning and trajectory control in reaching from an origin to destination. At intermediate levels of automation (semiautonomous), an intelligent vehicle acts as a copilot or driver assistant to guide the driver in many aspects of driving seamlessly and as needed. As a pure driving aid, an intelligent vehicle acts as an informational assistant to alert, warn, or caution the driver, while leaving the all actual controls to the driver. On the other hand, some active

safety functions of intelligent vehicles that handle the vehicle dynamics are achieved totally automatically and transparent to the driver. In this category, ABS, traction control, and electronic stability control, among other systems, sense the situation at hand and the dynamics of the relevant vehicle components and actuate as necessary to improve vehicle stability and safety.

6 Purpose and Scope of This Handbook

6.1 Purpose

As mentioned earlier, the main goals of intelligent vehicles are to increase safety and comfort of driving, while improving efficiency and reducing environmental impact. Significant progress has been made toward these goals by developing active safety systems, which are progressively incorporating more intelligence in the vehicles. As opposed to passive safety systems, which are inherent in the optimal structural design to mitigate crashes (e.g., vehicle crush zone, safety belts, head restraints, etc.), the active safety systems engage preemptively to actuate vehicle systems to either avoid a collision or reduce the severity of impacts (e.g., traction control, electronic stability control, brake assist, etc.) The progress has been motivated commercially by the industry and reinforced publicly by the governments to improve vehicle safety. To expand the range and scope of active safety systems, intelligent vehicles provide information, warnings and alerts in addition to supporting the primary tasks of controlling the vehicle and the secondary tasks of manipulating in-vehicle devices (radio, phone, navigation, etc.).

It is safe to say that the primary goal of present intelligent vehicles is to improve the safety and comfort of driving. However, in the long term, such vehicles are expected to provide total trip planning and execution and to give the option of relieving the driver from the driving task during an entire trip while ensuring full safety and the most efficient travel choices.

There has been enormous progress in the development of intelligent vehicles from both industry and government. However, the science behind most of the industrial development has been proprietary and remains unpublished. Government efforts have focused on exploratory analyses, identification of requirements, development of standards, and laboratory and field operational tests and evaluations. The technical methods and engineering of intelligent vehicles have been disseminated through a number of scientific and engineering journals and conferences such as those sponsored by IEEE ITS Society ([IEEE ITS Web site 2011](#)), ITS America, SAE, and others. Even earlier, there had been valuable efforts by researchers to report findings in intelligent vehicles, either from a project perspective or with a focus on specific aspects like navigation, etc. (Bishop 2005; Vlacic et al. 2001; Catling 1993; Michon 1993). However, as the progress and development has been rapid, there is no current source to culminate technological developments and new emerging systems in a comprehensive format.

This handbook provides an up-to-date review of technologies and methods that are used for intelligent vehicles. The focus of this handbook is on road vehicles, that is,

passenger cars, SUVs, minivans, trucks, and buses, as they all share similar technologies. It provides case studies of prototypes and existing commercial systems and highlights many research areas and future challenges. Some chapters focus on engineering methods while others emphasize systems functionality and requirements. All chapters provide a review of the state-of-the-art in respective areas where applicable, and present the progress made to date. All chapters provide a rich bibliography that readers can refer to for additional material. This handbook covers both the important fundamentals and the latest technologies and methods that an engineer or manager in this field needs to know about intelligent vehicles.

Eleven parts containing multiple chapters are organized in an easy-to-follow format to introduce the major systems, technical methods, and issues of intelligent vehicles that deal with the vehicle, the human, and their interactions. Due to the interdisciplinary nature of the field and the interactions between vehicle systems and humans (driver), there are inevitably some subject overlaps among different parts of this handbook. For example, in the coverage of driver assistance, while fundamentals of interaction between the vehicle and driver are covered, case studies present safety systems that aid the driver. On the other hand, the coverage of vehicular safety also details the various available in-vehicle safety systems. Furthermore, when covering vision processing, inevitably, some vision-based safety systems are reviewed again. Thus, in some cases, multiple chapters, although having different goals, perspectives, and coverage, may discuss a common system or topic. While the objectives of each part (and the corresponding chapters) are different, these partial overlaps among topics provide not only additional coverage but also various views on the subject; they further emphasize the interrelationship between the vehicle and driver as viewed by experts in their respective technical fields.

6.2 Coverage of Topics

This *introductory chapter* provides the scope of the handbook, an introduction to the significance of the intelligent vehicles, and a brief historical perspective. The section on *Overview of Intelligent Vehicle Systems and Approaches* provides a detailed analysis of intelligent vehicle systems and approaches; its chapters cover a range of topics from fundamental strategies to behavioral adaptations and simulation methods.

Since handling a vehicle trajectory automatically is a major contribution of intelligent vehicles, the section on *Vehicle Longitudinal and Lateral Control Systems* covers methods of longitudinal and lateral control of vehicles as well as coordinated control of multiple vehicles from a vehicle dynamics perspective. In-depth formulations and control design methods are presented in each chapter. Case studies in each chapter reflect the successful implementations of the presented methods.

Special Vehicular Systems covers two distinct important vehicular systems, one covering drive-by-wire systems and another, vehicular energy systems. As the need for automation in intelligent vehicles increases, mechanical actuation (exerted by driver direct control) needs to be replaced by its electronic and electromechanically actuated

counterparts to enable automatic braking, acceleration, steering, etc. Therefore, one chapter introduces the design and functionalities of drive-by-wire systems. Although not solely an intelligent vehicle topic, energy systems are a major component in the development of smart and efficient vehicles of the future, and hence, a chapter is dedicated to this very important topic.

Positioning, Navigation, and Trajectory Control covers vehicle guidance and trajectory control from a global positioning perspective. It covers topics in satellite systems, mapping, vehicle positioning, and in-car navigation.

Driver Assistance focuses on methods of providing assistance to the driver with a fundamental look into driver requirements in addition to modeling and testing methods. As an example of an extensively tested driver assistance system with significant safety implications, a whole chapter is dedicated to intelligent speed adaptation (ISA) systems.

Safety and Comfort Systems covers a comprehensive list of major vehicular safety and comfort systems over 11 chapters. The coverage of each chapter includes the intended use of the systems, design of their components, and their implementations. Most of these systems are either commercially available as options in new vehicles or have been developed in prototypes.

Drowsy and Fatigued Driver Detection, Monitoring, Warning addresses the more specific and a very challenging problem of assisting fatigued and drowsy drivers. Due to significant variability in driver conditions, reactions, and responses under these strenuous conditions, detection and development of countermeasures has been exceptionally difficult. The research areas involved in detecting fatigue and drowsiness and development of effective countermeasures, and existing prototypes and commercial systems are covered in four chapters.

A main enabling feature for intelligent vehicles has been new sensing and situational awareness capabilities afforded by advanced sensors. The crown jewels of these sensing capabilities are vision and radars. Since vision has been successfully implemented in many safety systems, the entirety of *Vision-Based Systems* is dedicated to this topic. Seven chapters cover a range of topics spanning from image processing and camera technology to various perception capabilities and safety features made possible by vision systems.

Vehicular Communications Systems covers one of the most important emerging technologies in the automotive and transportation industry, namely, wireless communications. Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications are anticipated to reshape the future of driving by significantly enhancing coordinated driving and enabling various critical safety systems. Five chapters cover the main issues involved in these new developments.

Fully Autonomous Driving covers the ultimate level of automation in intelligent vehicles, that is, fully autonomous (driverless) driving. Autonomous driving can be considered the ultimate driver assistance, totally relieving the driver from the driving task. Prototypes have been developed and several competitions worldwide have demonstrated autonomous driving. Despite the successful demonstration of the feasibility of autonomous driving, its full implementations in real traffic and road environments with

unanticipated disturbances still face several safety challenges. Eight chapters discuss, in depth, the various elements required for autonomous driving and its challenges.

Finally, the section on *A Look to the Future of Intelligent Vehicles* provides a perspective on the future of intelligent vehicles while addressing some of the main implementation challenges as well as the market outlook for such technologies.

Note that each part and chapter, while related to the comprehensive scope of the handbook, is written in an independent fashion intended to be useful for readers of all backgrounds, whether interested only in one or in many specific topics.

Another important note is that this handbook is totally commercial free. A proper coverage of the available intelligent vehicle systems necessitates occasional mention of various OEMs (automobile manufacturers) and their respective optional systems or prototypes where applicable. However, in coverage of existing systems and case studies, examples drawn from various OEMs should not be construed in any way a commercial endorsement or preference given to any manufacturer or respective system. The presented systems are merely examples, selected from many available ones, for illustrative purposes to ensure an appropriate coverage of the state-of-the-technology or state-of-the-art.

Finally, the handbook is a comprehensive and up-to-date resource for the study, analysis, design, and development of the intelligent vehicles. It is intended to serve professionals, engineers, scientists, and students from industry, government, academia, and general public in their own research and independent investigation of the presented topics.

References

-
- Bishop R (2005) Intelligent vehicles technology and trends. Artech House, Norwood
- Catling I (ed) (1993) Advanced technology for road transport: IVHS and ATT. Artech House, Boston
- DOT HS (2010) Traffic safety facts research note: summary of statistical findings. Highlights of 2009 motor vehicle crashes, DOT HS 811 363, Washington, DC. <http://www-nrd.nhtsa.dot.gov/cats/index.aspx>
- DOT HS, US DOT NHTSA (2009) Traffic safety facts a compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system, Early edn. DOT HS 811 402, US DOT NHTSA, Washington, DC. <http://www-nrd.nhtsa.dot.gov/CATS/>
- EnergyIndependence. <http://www.americanenergyindependence.com/fuels.aspx>
- <http://www.merriam-webster.com/dictionary/intelligent>
- Intelligent Transportation Society of IEEE. <http://ewh.ieee.org/tc/its/>
- ITS JPO DoT. Intelligent Transportation Systems (ITS) program overview. http://www.its.dot.gov/its_program/about_its.htm
- Michon JA (ed) (1993) Generic intelligent driver support. Taylor and Francis, London
- National Transportation Statistics, US DOT Research and Innovative Technology Administration, Bureau of Statistics, Quarterly Publications. http://www.bts.gov/publications/national_transportation_statistics/
- Swedish Government Health and Welfare Statistical Database (2006) <http://192.137.163.40/epcfs/index.asp?kod=engelska>
- Texas Transportation Institute (1990) Proceedings of a national workshop on IVHS sponsored by Mobility 2000, Dallas. Texas Transportation Institute, College Station. http://ntl.bts.gov/lib/jpodocs/repts_te/9063.pdf
- US Energy Information administration. Use of energy in the United States explained. http://www.eia.gov/energyexplained/index.cfm?page=us_energy_use

Vlacic L, Parent M, Harashima F (2001) Intelligent vehicle technologies, Society of Automotive Engineers (SAE) international. Butterworth-Heinemann, Boston

World Health Organization (2009) Global status report on road safety: time for action. World Health

Organization, Geneva. www.who.int/violence_injury_prevention/road_safety_status/2009.

Zhao Y (1997) Vehicle location and navigation systems. Artech House, Boston

Overview of Intelligent Vehicle Systems and Approaches

Bart Van Arem

2 A Strategic Approach to Intelligent Functions in Vehicles

Bart van Arem

Civil Engineering and Geoscience Transport & Planning, Delft
University of Technology, Delft, CN, The Netherlands

1	<i>Introduction</i>	18
2	<i>Classification</i>	19
3	<i>Route Guidance Systems</i>	21
4	<i>Advanced Driver Assistance</i>	23
5	<i>Automatic Vehicles</i>	25
6	<i>Non-technological Issues</i>	26
7	<i>Conclusions</i>	28

Abstract: Intelligent vehicles can make road traffic safer, more efficient, and cleaner. We provide an overview of current and emerging intelligent functions in vehicles. We focus on intelligent functions that actively interfere with the task of driving a vehicle. We give a classification based on the driving task, the type of road, and the level of support. We give a review of route guidance systems, advanced driver assistance systems, as well as automated vehicles. We discuss the non-technological issues that need to be addressed for the successful deployment of intelligent vehicles, such as cooperation between industry and public authorities, increasing awareness amongst stakeholders, research and development, and legal framework. Finally, we introduce the subjects that will be addressed in the remainder of this section.

1 Introduction

Human error contributes to most of the road accidents. Human drivers have limitations in terms of reaction time, perception, and control. Factors such as high workload and fatigue can aggravate these limitations. Can new technologies such as sensing and communication equipment lead to intelligent functions to drive more safely and efficiently? On the other hand, human drivers have the capability of understanding complex situations that they have not experienced before and to identify intentions of other road users from subtle clues. Could new technology ever reach this level of sophistication?

Automated driving has been an element in many future visions of mobility, enabling safe, efficient, reliable, and clean transportation. During the last decades, technological progress in the field of sensing, communication, and information processing has led to an increasing interest in intelligent functions in vehicles. This interest does not only pertain to automated driving. It also pertains to intelligent functions that support a driver while the driver is assumed to remain in control. An example of such a system is Adaptive Cruise Control (ACC): ACC controls the speed and headway of the vehicle, but it can be overruled by the driver, and it even must be overruled by the driver in the case of an imminent collision that is beyond the operational scope of the ACC.

Governments and road operators throughout the world show an increasing interest in intelligent vehicles because they expect that intelligent vehicles can make road traffic safer, more efficient, and cleaner. Traffic safety can be improved because intelligent functions in vehicles can help a driver anticipate dangerous conditions, for instance, by warning a driver for dangerous conditions such as imminent collisions. Traffic efficiency can be improved by functions that help a driver to avoid congestion or by reducing strong accelerations and decelerations or (safely) maintaining shorter headways. Reducing congestion and smoothening acceleration and deceleration behavior will also result in reduced emissions. In addition, intelligent functions could help avoid driving through areas with a high vulnerability to emissions.

The aim of this chapter is to provide an overview of intelligent vehicles from a functional point of view. It will provide a methodology for classifying intelligent functions in vehicles and give a review of route guidance and advanced driver assistance

systems as well as automatic vehicles. Next, it will discuss supporting actions that are needed for large-scale deployment of intelligent vehicles. This chapter will conclude with an outlook.

2 Classification

As there are many different intelligent functions, it is useful to classify them. In the chapter, we focus on intelligent functions that actively interfere with the task of driving a vehicle. We give a classification based on the driving task, the type of road, and the level of support.

The task of driving a vehicle actually consists of different interacting subtasks. We give a classification based on the notion that the driving task can be seen as a three-layered hierarchical task with a strategic, a tactical, and an operational level (► [Table 2.1](#)) (Michon 1985). The highest, strategic level is also called the navigation level. At this level, the goals of the trip are set in terms of destination and route. The strategic level can also involve the way the driving style, such as hurried, fuel efficient, enjoying the ride, cautious. The typical time scale on which strategic goals are (re-)evaluated is in the order of minutes or more. Navigation systems are intelligent functions that support the strategic driving task. The second level is also called the maneuvering level. At this level, interaction with other road users and the road layout takes place. Examples are overtaking a vehicle, car following, or negotiating an intersection. The time scale for the tactical driving tasks is in the order of 10 s. Blind spot warning systems intelligent functions that support the tactical driving task. The lowest level – operational level or the control level – describes the actual control of the vehicles, including steering, throttle, brake, and clutch. The time scale is in the order 1 s or smaller, or continuous. Intelligent functions that support the driver at the operational level are Electronic Stability Control (ESC) or (Adaptive) Cruise Control.

Intelligent functions may function very differently in different types of traffic. ► [Table 2.2](#) summarizes general characteristics for motorway, rural, and urban traffic. Motorway traffic is least complex, amongst other by the one way driving and standardized geometry. Therefore, intelligent functions may be introduced for motorway traffic, for example, Adaptive Cruise Control, which controls the speed of a vehicle while

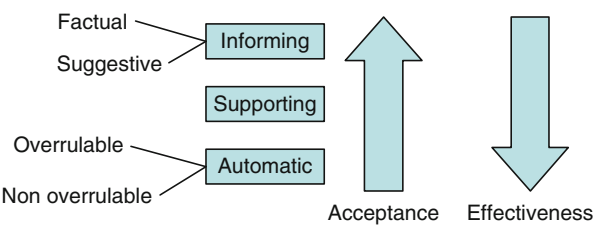
■ **Table 2.1**

Examples of driving task at hierarchical levels

Level of driving task	Examples
Strategic – navigation	Destination, route, driving style
Tactical – maneuvering	Lane changing, turning
Operational – control	Steering, throttle, brake

■ Table 2.2
General characteristics for motorway, rural, and urban traffic

Motorway traffic	Rural traffic	Urban traffic
Most homogeneous class of roads with rather uniform traffic conditions	Non-motorway connections outside urban built-up areas	Heterogeneous class, widely different in terms of size and density
Sparse network, few intersections and entry/egress points, traffic separated by direction	Moderately dense network, two-way traffic, limited intersections with traffic control	Dense and complex road networks, intersections and crossings; two-way traffic, traffic signals
Moderate to very high traffic flow levels, homogeneous traffic, generally rather high speeds except in congestion with stop-and-go conditions, standardized and predictable road geometry	Mixed traffic but mainly used by motorized traffic, wide range of driving speeds, wide variety of road geometry	Varying traffic loads, complex traffic composition
Low to moderate levels of driver attention, except in (nearly) congested conditions	Moderate to high driver loads	Low to moderate speeds, heavy driver load



■ Fig. 2.1
Level of support of intelligent functions

safeguarding a minimal distance to a predecessor. Rural traffic is more complex than motorway traffic, amongst others because of two-way traffic and intersections. Typical intelligent functions for rural traffic are lane keeping assistance and speed warning in curves. Urban traffic is most complex, having a large variety of intersection and road users. Typical intelligent functions for urban traffic are intersection collision warning and parking guidance.

The final classification is the level of support of intelligent functions. We distinguish informing, supporting, and automatic functions, see ● Fig. 2.1.

Informing system provides information to the driver but leaves the decision to act based upon the information fully to the driver. Information can be given through in visual format and/or through sound. Visual information can be provided through an

information display (text or icons), signaling lights or head-up display. Visual information has the advantage that it can be presented during a longer period without disturbing the driver. On the other, it may lead to distraction, depending on the position where the information is presented and the ease with which the information can be understood. Information can also be conveyed through sound such as speech and/or warning signals. The use of sound can convey a message effectively but cannot be repeated without becoming obtrusive to the driver. The content of the information may be factual “congestion in 500 m” or suggest a course of actions “turn right,” “reduce speed,” or “watch out for slippery road.”

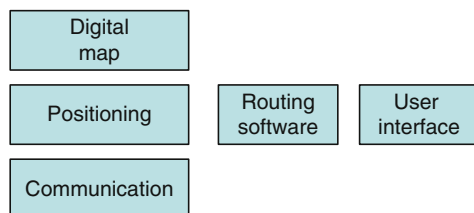
Supporting functions provide active support to the driver but leave the driver in control. They are usually based on subtle haptic cues that initiate the right course of action. Examples are lane keeping support, where a small force on the steering wheel is applied if the vehicle is unintentionally drifting out of its lane or an active accelerator pedal which gives counterforce on the accelerator pedal to urge the driver to reduce speed. Importantly, the driver is able to override the systems but needs to exert some additional force to do so.

In automatic systems, certain driving tasks are carried out by intelligent functions. Automatic intelligent functions can be either overrutable or non-overrutable. An overrutable system can be switched on or switched off by the driver. An example is Adaptive Cruise Control, which automatically controls the speed and headway of a vehicle. However, the driver can switch it on and off and can often also choose headway and speed setting. In situations outside the operational range of the ACC such as strong braking, the driver has to override the system. Emergency Braking is an example of a non-overrutable system, because in critical situations, the system needs to intervene directly without the delay of driver activation or confirmation. Other examples are fully automated systems, that have no driver at all or platooning systems which operates at short headways beyond human capabilities.

Evidently, the level of support is decisive for user acceptance and the effectiveness of the system. Typically, informing systems are expected to be appreciated by user because of their non-obtrusive nature and for leaving freedom to act to the driver (van Driel and van Arem 2005). Supporting and automated systems may be the most effective because of their active nature. Although drivers tend to take a skeptical position toward supporting and automatic systems, the acceptance increases after gaining experience and trust in these systems (van Driel et al. 2007).

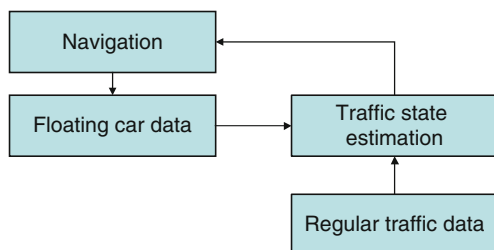
3 Route Guidance Systems

Route guidance systems are systems that support a driver in planning and following a suitable – usually the fastest – route to a destination. It is an informing system at the strategic and tactical level of the driving task. The technical basis of operation consists of a digital map, a position system such as GPS, routing software, a user interface and, increasingly, communication systems (► Fig. 2.2).



■ Fig. 2.2

Technical basis for route guidance systems



■ Fig. 2.3

Real-time traffic information using floating car data

The digital map contains information about the network and road topology and geometry. The positioning system, usually a GPS, provides the position in geographical coordinates which is subsequently used for the localization of the vehicle on the digital map. The routing software can be considered as the kernel of the system. It uses mathematical algorithms such as Dijkstra's shortest path algorithm to compute the optimal route to the destination. The user interface is based on a touch screen display and sound and is used for user input as well as system output. Increasingly, route guidance systems are being equipped with communication systems based on cellular communication such as GPRS and UMTS.

During the past 10 years, navigation systems have become widely available. The first commercially available systems were integrated in the dashboard of top class vehicles. Positioning was done by a combination of GPS and vehicle data. The availability of a more accurate GPS signal up to 10 m opened up the possibility of running navigation software on Personal Digital Assistants and Smartphones.

The access of a navigation function to communication functions can be used for generating floating car data and for accessing remote services, such as fuel prices, alerts for speed cameras, etc. Of particular interest from the point of view of traffic safety and traffic efficiency is the possibility to obtain real-time traffic information, see ● Fig. 2.3.

The navigation function and its communication function are used to generate using so-called floating car data, consisting of regular time-stamped data such as the position, speed, friction, and status of wipers of the equipped vehicles. These data are used in combination with regular traffic state estimates collected by road providers to generate

a traffic state estimate. The traffic state estimation is used to provide information to the navigation system, such as warnings for dangerous situations (fog, slipperiness) or information about congestion on the road network. Modeling studies have shown that real-time traffic condition can result in 30% time savings in the case of road network which are congested due to an accident (Lee and Park 2008).

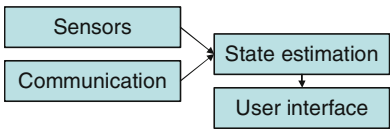
4 Advanced Driver Assistance

Advanced driver assistance systems support the tactical and operational driving task. The technical basis (● Fig. 2.4) consists of sensors that are used to estimate the state of the vehicle in relation to its direct environment and a user interface. In the future, short range communication systems can be used to obtain additional data from other vehicles and/or roadside systems.

Typical sensor systems used for ADA systems are radar systems, vision-based systems, infrared sensors, laser scanner systems. Each of these systems has its specific advantages and disadvantages, and they may also be applied in combination. The common element of the sensor systems is that they are aimed at providing an image of the direct surroundings of a vehicle in terms of other vehicles, road markings, other road users, road signs, etc. The typical range of outward-looking sensors is some 100 m. Inward-looking sensor systems could also be used to detect the state of a driver in terms of drowsiness or distraction. In addition, high-resolution data from digital maps in combination with accurate positioning can serve as an additional source of data.

In the future, short-range communication systems are expected to provide additional data to a vehicle. A special WLAN IEEE 802.11p has been developed for communication between vehicles and between vehicles and roadside systems. This enables vehicles to exchange data on their status such as position, speed, heading, acceleration, outside temperature, frictions, and windscreen wipers. Roadside communication systems can store and forward information from vehicles or send information to vehicles such as the status of traffic lights, speed advices, restrictions, etc. Compared with GPRS/UMTS used in navigation systems, WLAN is a direct and fast way of communication on close distances up to some 100 m.

Next, the sensor data are used to estimate the state of the vehicle in relation to its context and for deciding on a course of action to be taken by the driver and/or vehicle. This is illustrated by a number of examples:



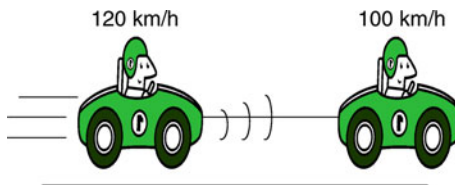
■ Fig. 2.4
Technical basis for advanced driver assistance systems

- A combination of detection of steering movements from the steering sensor and eye blinks from a camera is used to diagnose whether a driver is drowsy; the course of action is to alert the driver and/or suggest to take a break.
- Data from a forward looking radar and the speedometer is used to decide whether the vehicle is following its predecessor too closely; the course of action is to increase the headway.
- Data from a camera for pedestrian detection at the roadside is communicated to the vehicle, resulting in an imminent collision of the vehicle with the pedestrian; the course of action is to brake or steer to avoid the collision.
- Data from a camera is used to assess the presence of another vehicle in a vehicle's blind spot; the course of action is to stay in the present lane while the other vehicle is in the blind spot.

Finally, the function of the user interface is to convey the course of action to be taken to the user. Usually, a combination of user interfacing options (informing, supporting, automatic) is used.

The commercial introduction of Advanced Driver Assistance systems started in the late 1990s with the introduction of Adaptive Cruise Control (ACC) on top models of luxury brands. Nowadays, ACC is also available on medium-class car models (► Fig. 2.5). Most models are equipped with a radar sensor used to assess the speed difference and space headway to the predecessor. The system automatically adapts the vehicle's speed if the space headway is too small. The driver can switch the ACC on and off and can set the speed and headway of the system (within certain boundaries). The first-generation ACC systems operated only at speeds above 50 km/h and did not actively brake. Nowadays, ACC systems are also capable of operating at speeds up to stand still and apply the brake actively. However, emergency braking is outside of the operational range of the ACC. In such cases, the driver needs to intervene. A modeling study shows that congestion delay can decrease by 30% if 10% of the vehicles are equipped with ACC (van Driel and van Arem 2010).

Other Advanced Driver Assistance that are commercially available nowadays are Electronic Stability Control (ESC), Lane Keeping Support (LKS), Parking assistance systems, blind spot warning, and driver state monitoring. Systems based on vehicle–vehicle and vehicle–roadside communication have been studied in numerous research projects, but their commercial introduction is expected around 2015–2020.



■ Fig. 2.5
Adaptive cruise control

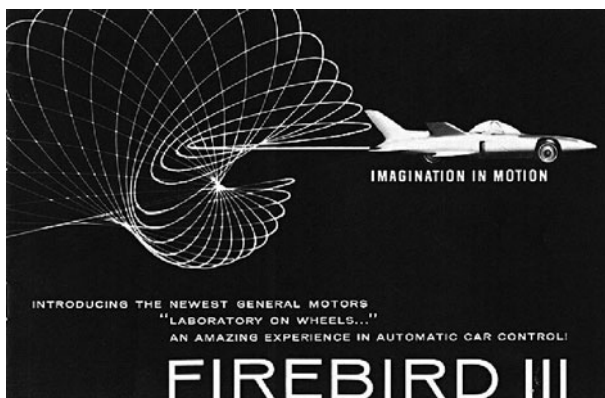
5 Automatic Vehicles

The vision of automatic driving has appealed to many researchers. Already in the 1950s, rocket science inspired automotive engineers to study jet-propelled automated vehicles (► Fig. 2.6).

Fully automated driving systems completely replace the human driver. The only task of the human “operator” is to switch on the system and to specify the destination. The advantage of fully automated driving is that the system can perform actions such as sustained driving at short headways in platoons of vehicles that are beyond the capabilities of human drivers. In order to maintain the appropriate safety levels, the requirements with respect to system safety in terms of failure rate, robustness, and graceful degradation are very high. In the late 1990s, several research programs throughout the world were conducted into automated motorways. The PATH program (► Fig. 2.7) developed automated driving in platoons of vehicles using a combination of magnet embedded in the road, on board radar systems, and vehicle–vehicle communication. Other research projects relied on vision and enhanced GPS.

It is commonly agreed that automated driving on motorways requires a dedicated, properly equipped road infrastructure. In addition to the technological challenges, this raises the challenge of the design of a transition from conventional to automated motorways. In addition, there are non-technological challenges such as legal issues and user acceptance.

There are good possibilities to apply automated vehicles on road infrastructure with controlled access. Robot vehicles are well known and successful in industrial automation. In the port of Rotterdam, automated vehicles are used to move containers on the terminal, using a fixed grid of transponders and a centralized control system. At several locations in the Netherlands, automated road vehicles have been used as shuttle services in connection to public transport facilities and parking lots. The automated people movers operate on



■ Fig. 2.6

Firebird program by general motors in 1959



■ Fig. 2.7

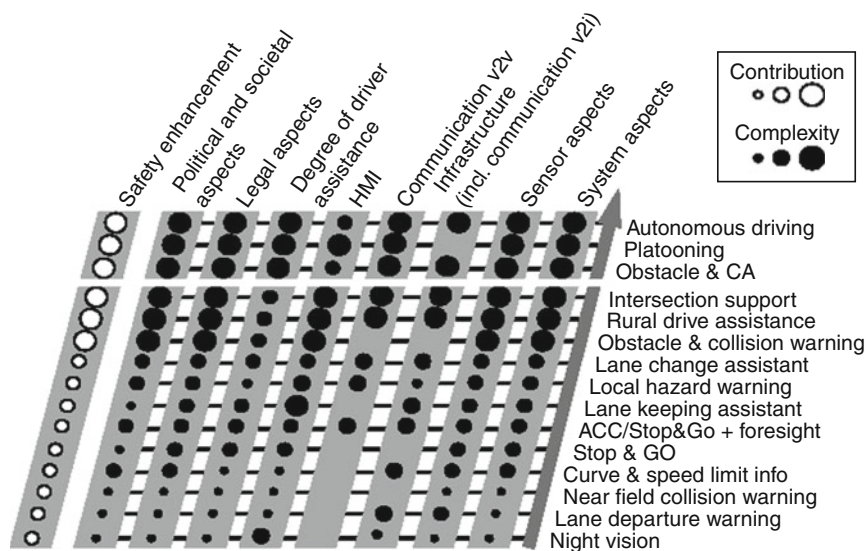
Automated driving by the PATH program in the Netherlands in 1997

a dedicated road equipped with transponders. The vehicles are equipped with sensors to detect obstacles and communication with a central control center.

Finally, automated vehicles can contribute to the safety of troops in military operations by applying driverless operations in potentially hostile environments. The US Defense Advanced Research Projects Agency (DARPA) has organized a series of challenges in automated driving in which teams of universities and industry competed to complete a specially design course. The first two challenges were aimed at completing a 200 miles drive in desert terrain, and the third was aimed at driverless operation of multiple vehicles in an urban environment. Most vehicles used GPS localization, digital maps, and different types of sensors such as vision and laser scanner to characterize the environment. Especially in the urban challenge, advanced algorithms were applied for trajectory planning and maneuvering.

6 Non-technological Issues

Evidently, intelligent vehicles are based on the results of technological research. In order to move from research to the eventual deployment of intelligent vehicles, also, non-technological issues need to be addressed. In the EU project ADASE 2 (Ehmanns and Spannheimer 2003), a road map of intelligent vehicles was developed jointly by the industry and road operators (► Fig. 2.8).



■ Fig. 2.8
Road map for advanced driver assistance systems

The road map shows different applications in increasing complexity, their expected contribution to traffic safety, political and societal aspects, and legal aspects.

The contribution to traffic safety was estimated based by consulting traffic safety experts. Although the potential contribution of intelligent vehicles to traffic safety is high, it must be taken into account that these benefits may be compensated by secondary impacts such as behavioral adaptation. Still, there is an expectation that intelligent vehicles could reduce the number of road fatalities by some 25% (see Kulmala 2010).

The political and societal aspects refer to the level of cooperation needed to develop and introduce intelligent vehicles. Systems such as lane departure warning can be introduced by the industry, requiring a minimal level of cooperation in the field of standardization. The potential of intelligent vehicles has been recognized by public authorities as well. At the EU level, three main pillars have been identified to promote the successful deployment of intelligent vehicles. The first pillar is cooperation between different industry sectors, public authorities, and road operators, which has resulted in the eSafety Forum, further resulting in the ITS Action plan that has been endorsed by the European parliament. The second pillar is aimed at improving the awareness of the industry, public authorities, and end users of intelligent vehicles. It has resulted in the EU-wide “Choose ESC” campaign as well as the inclusion of intelligent vehicle function in the EuroNCAP vehicle safety classification. The third pillar is aimed at R&D into intelligent vehicles, especially aimed at large-scale field operational test in order to understand the technical behavior of intelligent vehicles when applied at a large scale and to understand the user acceptance of and behavioral response to intelligent vehicles. Increasingly, the research is

aimed at cooperative systems using vehicle–vehicle and vehicle–roadside communication, especially addressing the communication networks and state estimation on the basis of distributed data. Finally, the research is aimed at a more profound understanding of the contribution of intelligent vehicles to safer, cleaner, and more efficient traffic. In the United States and Japan, similar actions are undertaken.

The legal aspects refer to the changes in the legal framework needed for the introduction of intelligent vehicles. An obvious issue is associated with the liability of suppliers of intelligent vehicles in case of malfunctioning. Normally, a driver is assumed to be in full control of the vehicle. If driving tasks are taken over by an intelligent function, the question is raised whether a driver can still be assumed to be fully responsible for the control of the vehicle, especially in the case of supporting or automatic driving support functions. The issue of liability is dealt with differently in different regions of the world. In the EU, a Code of Practice was developed (RESPONSE 3 2006) to provide the vehicle industry with the tools and common understanding to overcome and to help managing the problems about safety concerns and liability of Advanced Driver Assistance Systems. The application of the Code of Practice is a possibility to demonstrate that state-of-the-art procedures in ADAS development have been applied, including risk identification, risk assessment, and evaluation methodology. Other legal issues to be taken into account are the legal tasks and responsibilities of public authorities and road operators. Legal frameworks with respect to privacy and security are relevant, especially in the case of information exchange using vehicle–vehicle and vehicle–roadside communication.

7 Conclusions

Vehicles are increasingly becoming intelligent by the application of intelligent functions. This trend is driven by the increased possibilities of information and communication technology, sensors, and computing on the one hand, as well as the expectation that intelligent functions can lead to safer, cleaner, and more efficient traffic. At the strategic level of the driving task, route guidance systems are widely used nowadays and offer a good basis for the provision of additional services such as traffic information. At the tactical and operational level of the driving task, advanced driver assistance systems are slowly gaining momentum. Automated vehicles are applied in markets such as people movers, industrial automation, and unmanned military operations. The successful deployment of intelligent vehicles also depends on effective cooperation between the industry, road operators, and public authorities in order to jointly address legal barriers, standardization, increasing awareness amongst end user, and initiating joint research programs, especially on field operational tests, vehicle–vehicle and vehicle–roadside communication, and the study of user acceptance and behavioral adaptation and impact on traffic flows.

In the next chapters of this section “Overview of Intelligent Vehicle Systems and Approaches,” the following key issues will be elaborated. Chapter 2 will focus on sensors and actuation, providing an overview of the increasing possibilities of sensors and

actuators for intelligent vehicles. ➤ **Chapter 3** will also include vehicle communication for situational awareness of a vehicle's surroundings. In ➤ **Chap. 4**, we focus on algorithms for hierarchical, intelligent, and automatic control of intelligent vehicles. ➤ **Chapter 5** addresses the key issue of behavioral adaptation and driver acceptance. We conclude in ➤ **Chap. 6** by reviewing modeling approaches for the design and evaluation of intelligent vehicles.

References

- Ehmanns D, Spannheimer H (2003) Roadmap, Deliverable D2D of the European project ADASE- II (IST-2000-28010)
- Kulmala R (2010) Ex-ante assessment of the safety effects of intelligent transport systems. *Accid Anal Prev* 42:1359–1369. doi:10.1016/j.aap. 2010.03.001
- Lee J, Park B (2008) Evaluation of vehicle infrastructure integration (VII) based route guidance strategies under incident conditions. In: *Proceedings of the 87th TRB 2008 annual meeting*, Washington, DC
- Michon JA (1985) A critical view of driver behavior models: What do we know, what should we do? In: Evans L, Schwing RC (eds) *Human behavior and traffic safety*. Plenum, New York
- RESPONSE 3 (2006) D11.2 Code of practice for the design and evaluation of ADAS, Deliverable RESPONSE3 subproject V3.0, 31.10.2006. www.prevent-ip.org, Accessed 27 Dec 2010
- van Driel CJG, Hoedemaeker M, van Arem B (2007) Impacts of a congestion assistant on driving behaviour and acceptance using a driving simulator. *Transp Res F* 10(2):139–152
- van Driel CJG, van Arem B (2005) Investigation of user needs for driver assistance: results of an Internet questionnaire. *Eur J Transp Infrastruct Res* 5(4):297–316
- van Driel CJG, van Arem B (2010) The impact of a congestion assistant on traffic flow efficiency and safety in congested traffic caused by a lane drop. *J Intell Transp Syst: Technol Plann Operat* 14(4):197–208

3 Sensing and Actuation in Intelligent Vehicles

Angelos Amditis¹ · Panagiotis Lytrivis¹ · Evangelia Portouli²

¹Institute of Communication and Computer Systems (ICCS),
Zografou, Athens, Greece

²National Technical University of Athens, Zografou,
Athens, Greece

1	Introduction	33
2	Sensing	34
2.1	General In-Vehicle Sensors	34
2.1.1	Yaw Rate Sensor	34
2.1.2	Accelerometer	35
2.1.3	Wheel Speed Sensor	35
2.1.4	Steering Angle Sensor	35
2.1.5	Other Sensors	36
2.2	Perception Sensors	36
2.2.1	Radar Sensors	36
2.2.2	Laser Scanners	38
2.2.3	Vision Systems	39
2.2.4	Ultrasonic Sensors	40
2.3	Virtual Sensors	41
2.3.1	Digital Map	41
2.3.2	Wireless Communication	42
3	Actuation	44
3.1	Categories of Actuators According to Energy Source	45
3.1.1	Mechanical Actuators	45
3.1.2	Electrical Actuators	45
3.1.3	Pneumatic and Hydraulic Actuators	48
3.1.4	Piezoelectric Actuators	49
3.1.5	Thermal Bimorphs	50
3.2	ABS, ESC and ACC Systems	50

3.3	Assisted Steering and Steer-by-Wire Systems	52
3.4	Brake-by-Wire Systems	53
3.5	Highly Automated Vehicles	54
3.5.1	Autonomous Vehicles Demonstrated in the Past	55
3.5.2	Advanced Automated Systems Under Development	55
4	Conclusion	58

Abstract: Nowadays, cars are equipped with more electronic systems than in the past. Today, vehicles are equipped with hundreds of miniature sensing systems, such as temperature, tire pressure, accelerometer, and speed sensors. Actuators are also installed as components of advanced systems for optimized braking and assisted steering, although the market for really intervening systems is not mature yet.


In this chapter, different sensing and actuation systems are highlighted. The purpose of this chapter is to give an overview of these systems and do not go deep into detail about the different technologies behind such systems. The sensors are grouped into three different categories: general in-vehicle sensors, perception sensors, and virtual sensors. Most of the general in-vehicle sensors are already available in the automotive market in the majority of commercial cars. On the other hand, the market penetration rate of perception sensors, except for ultrasonic sensors, is very low mainly because of their cost. Finally, there are some information sources that are not actual sensors and play a significant role in automotive applications, such as the digital maps. Actuators are first distinguished and described according to their energy source into mechanical actuators, electrical actuators, pneumatic or hydraulic actuators, piezoelectric actuators, and thermal bimorphs. Next, the design of advanced intervening systems is presented, namely, the ABS, electronic stability control, autonomous cruise control, assisted steering. More advanced systems like the steer by wire and brake by wire are also presented, although they have not yet entered the market as products, with few exceptions. Finally, the vision of a fully automated vehicle is presented together with the considerations that still accompany it, and some first prototypes and research work toward this direction are highlighted.

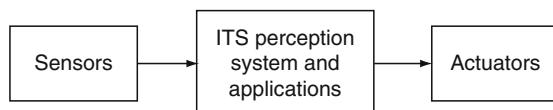
1 Introduction

Automotive electronics are now beginning to radically change the automobile in ways that would have seemed impossible just a few years ago. Engine control and ABS systems based on electronic subsystems are now standard features of most automobiles.

Vehicles today can *sense* the environment and *act* in case of emergency. Sensing the environment is more common than actuation in vehicles. The reason for that is that the sensing technology already counts some decades in the process of manufacturing the vehicles, whereas the use of actuators as parts of intervening systems has been recently introduced. Moreover, there are some other issues that limit the widespread usage of intervening systems, and thus of actuators. For example, if an accident occurs who should be responsible: the intervening system or the driver?

In general, the basic elements for the successful penetration of electronic systems such as sensors and actuators in the market are the miniaturization, the cost reduction, the increased functionality, and the quality of the component.

The input layer to an intelligent vehicle's system is the sensing layer which includes a variety of sensors, for example, GPS, radar, and lidars, while the output comprises the actuation layer. So sensing and actuation are closing the development loop of an intelligent vehicle's system, as shown in  Fig. 3.1.



■ Fig. 3.1

The concept of intelligent transportation systems

2 Sensing

Sensing the environment is one of the key elements of the future intelligent transportation systems. There is an enormous variety of different sensors used in the automotive industry: from in-vehicle sensors built-up together with the vehicle, perception sensors such as radars and lidars, as well as “virtual” sensors. With the word “virtual,” we are referring to information sources used widely in the automotive industry without being real sensors with the usual sense.

2.1 General In-Vehicle Sensors

This category includes the general sensors which are installed in the vehicle during its construction phase. There are hundreds of sensors that belong to this category, but here the focus will be on these sensors that are of interest for ADAS applications. The selection of the sensors that are briefly described in the following is based on the authors’ experience.

2.1.1 Yaw Rate Sensor

A yaw rate sensor measures the rotation of an object along a selected axis. The terms “gyroscope” and “angular rate sensor” are used alternatively in some cases. In the automotive field, a yaw rate sensor is used for measuring a vehicle’s angular velocity around its vertical axis. The unit for measuring the yaw rate is usually degrees per second ($^{\circ}/s$) or radians per second (rad/s).

The yaw rate sensors used in vehicles require particular attention. Coriolis accelerations in the range of mg must be detected correctly, and at the same time the accelerations occurring in the range of several g must not interfere with the sensor function.

The applications in the automotive sector that take advantage of this sensor are mainly the following:

- Electronic stability program (ESP) or vehicle dynamics control (VDC)
- Navigation (combined with GPS and accelerometer information)
- Rollover protection
- Curve speed warning (in combination with digital maps)

2.1.2 Accelerometer

An accelerometer is a device that measures proper acceleration, which is the acceleration it experiences relative to freefall. Both single-axis and multi-axis accelerometer models are available in the market. At this point, it should be highlighted that in some cases the yaw rate and accelerometer sensors are placed together in the same integrated circuit.

There is a significant amount of applications based on accelerometers, from medical and biology to gaming and navigation. In the automotive sector, the accelerometer sensor is used mainly for navigation purposes. An inertial navigation system (INS) is a navigation aid that uses a computer and motion sensors, such as accelerometers and yaw rate sensors, to continuously calculate via dead-reckoning the position, orientation, and velocity of a moving object without the need for external references.

2.1.3 Wheel Speed Sensor

A wheel speed sensor is a type of tachometer. It does not monitor vehicle speed directly but it senses the movement of the circumference of the tire. It actually reads the speed of a vehicle's wheel rotation. There are two main categories of wheel speed sensors: passive and active. Passive sensors do not need a power supply, whereas active sensors require an external power supply for operation.

Wheel speed sensors attached to wheels of a vehicle, respectively, to detect wheel speeds of a running vehicle are designed to detect the rotation of rotors which rotate together with axles coupled to the respective wheels. Wheel speed sensors for automobiles typically utilize an indexing disc mounted on a wheel and a pickup that detects the passage of marker elements carried by the disc as the wheel turns. This pickup can be mechanical, optical, or magnetic.

Wheel speed sensors are critical components of antilock braking systems (ABS), traction control systems (TCS), and similar functions.

2.1.4 Steering Angle Sensor

The overall steering wheel angle is measured by the steering angle sensor. The steering angle sensor is mounted on the steering shaft. Steering angle sensors were developed in the mid-1990s. Sensing a steering angle in automotive applications can be done in various ways using optical, magnetic, inductive, capacitive, or resistive sensor principles.

This sensor has two potentiometers offset by 90° . The steering wheel angle determined by these two potentiometers covers one full steering wheel turn; each of these values is repeated after $\pm 180^\circ$. The sensor knows this and counts the steering wheel revolutions accordingly. The overall steering wheel angle is thus made up of the current steering wheel angle together with the number of steering wheel rotations. In order that the overall steering wheel angle is available at any time, uninterrupted detection of all steering wheel movements – even when the vehicle is stationary – is required.

Steering angle sensors are used in cars for intelligent electronic control and assistance systems such as ESP, active steering, advanced front lighting, lane departure warning, and navigation.

2.1.5 Other Sensors

There are hundreds of other sensors that are used in a vehicle. Indicatively, some of them are tire pressure sensors, temperature sensors, rain sensors, and fuel sensors. For more information, the interested reader can refer to Marek et al. (2003).

2.2 Perception Sensors

Perception sensors are widely used for research purposes all around the world, but their penetration to the market is not so straightforward because of their cost which makes them a privilege of the luxury cars. Environmental perceptions systems for driver assistance and safety functions are based on radar, laser, vision, or ultrasonic sensors. These are the most important perception sensors, and their characteristics are analyzed in the following.

2.2.1 Radar Sensors

Radar technology is starting to be developed in the automotive industry mainly for the interests of road safety. There are two main categories of radar sensors designed for different purposes and tailored for different applications, namely, the *short-range radars* and the *long-range radars*. These radar categories have different technical characteristics and are operating in different frequencies. More details about the radar sensors are highlighted in the following paragraphs.

Short-Range Radar (SRR)

A “temporary” frequency band has been allocated at 24 GHz for the short-range radar sensors, allowing equipment to be marketed in the short term. However, this band is also used by other radio services that would suffer interference if too many radar devices were operated simultaneously in the same area. For this reason, this band will be closed for the introduction of new devices before the usage becomes too dense. A “permanent” band has been allocated at 79 GHz, allowing for long-term development of this radar service. European Commission Decision 2004/545/EC requires this band to be made available in all EU member states.

The object quantities measured are distance, velocity, and angle of detection. The SRR is an ultrawide band (UWB) pulse radar, which operates with a carrier frequency of 24 GHz. In principle, an ultrawide band pulse is transmitted, and the time of flight (TOF)

of the pulse from the sensor via an object return to the sensor is measured in order to calculate the distance to the reflecting object.

These sensors are mainly installed in the side areas of the vehicle and are used for detection of objects (other vehicles, guardrails, etc.) next to the vehicle. Moreover, they are used often for monitoring the rear area of the vehicle and sometimes they supplement the long-range radar sensor in the front area. Because the field of view of these sensors is limited, to cover the area of interest around the vehicle, usually they are installed as a network.

Anticollision is the main function relevant to an SRR sensor. This is being developed as part of a system to warn the driver of a pending collision, enabling avoiding action to be taken. In the event where collision is inevitable, the vehicle may prepare itself (e.g., by applying brakes, pretensioning seat belts) to minimize injury to passengers and others. Moreover, this sensor opens up the possibility of achieving the following functions:

- Parking aid
- Blind spot detection
- Precrash detection for front and side
- Short-range sensor for ACC stop and go function

Some general technical characteristics of the SRR sensor are given in [Table 3.1](#).

Long-Range Radar (LRR)

The operating frequency of the LRR sensors is typically in the 76–77 GHz band applying frequency-modulated continuous wave (FMCW) or pulse-Doppler operation.

The range of this sensor is up to 150–200 m and its field-of-view is only 11–12° (more details in [Table 3.2](#)). Usually, this sensor is located in the middle of the front bumper and it is used for frontal collision avoidance applications. Rarely this can also be found in the back bumper looking backward or two of them can be installed in the front bumper for extending the area covered in front.

■ Table 3.1

Short range radar specification

Detection range	0.2–50 m
Detection angle	±35°
Frequency	24 GHz (79 GHz)

■ Table 3.2

Long-range radar specification

Detection range	2–150 m
Detection angle	±6°
Frequency	76–77 GHz

In principle, this sensor detects metallic objects in the vehicle surrounding. It is able to detect multiple objects and to measure distance, relative speed, and the angle to an object simultaneously.

A typical application that is strongly related to the LRR sensor is adaptive cruise control (ACC) which enables the equipped vehicle to maintain a safe distance and speed from a vehicle in front. However, the LRR sensor can be used for other safety applications as well, such as collision avoidance.

Despite the fact that LRR utilizes weather-resistant technology, overall robustness against weather cannot be achieved. In extreme weather situations, full functionality of the sensor cannot be guaranteed. Large amounts of dirt may also reduce the sensor function. Moreover, topology limitations exist in curved road segments due to the narrow horizontal detection angle of this sensor.

2.2.2 Laser Scanners

Laser scanners take measurements according to the time-of-flight principle. A laser pulse with a defined duration is sent and reflected by an object. The reflection from the object is captured by a photodiode and transformed into signals in an optoelectronic circuit. The time interval between the pulse of light being sent and its reflection being received, making due allowance for the speed of light, indicates the distance to the object that reflected the light.

Laser scanners are generally robust, but have decreased sensitivity in adverse weather conditions. This fact limits their availability and reliability. Most common laser scanners provide range and bearing information with sub-degree resolution and accuracies on the order of 1–10 cm for 10–50 m ranges.

Laser scanners can be used for detecting other vehicles or obstacles in the road scene and vulnerable road users, such as pedestrians and cyclists. Moreover, the road can be detected by a laser scanner, especially the road borders.

Laser scanners exhibit much better lateral resolution as compared to radar sensors, but they come with comparatively slow scanning repetition rates, considerable physical size, and comparatively high production costs so far. Also they are significantly affected by the weather conditions.

The specification of different suppliers of laser scanners is not so identical. However, an indicative example of laser specifications is highlighted in [Table 3.3](#).

Both longitudinal and lateral vehicle control applications based on laser scanners exist but they are available only in expensive vehicles. Currently, the market penetration of these sensors is low and the penetration rate is slow.

Some example applications can be found in some European projects, for example, [PREVENT](#) and [SAFESPOT](#). There is also a European project called [MINIFAROS](#) with the objective to develop a miniature laser scanner which will be quite cheap with premium performance compared to off-the-shelf laser scanners and with fast and significant market penetration rate.

■ Table 3.3

Laser scanner specification

Detection range	min 0.3 m
	max 80 m (pedestrian, bike)
	120 m (motorbike)
	200 m (car, truck)
Field of view	Horizontal 150–240°
	Vertical ±2°

2.2.3 Vision Systems

Among the many new and emerging technologies for vehicular applications, automotive vision systems are a primary example. Indeed, video and vision processing have become the fastest growing technologies being deployed in vehicles.

This category comprises a variety of different vision systems and supported functions. These functions can range from lane recognition and traffic sign recognition to object detection (pedestrians, vehicles, etc.). The vision systems used in automotive applications can be categorized to:

- Charge-coupled device (CCD) sensors
- Complementary metal oxide semiconductor (CMOS) sensors
- Infrared (IR) sensors
- Stereovision systems

CCD and CMOS Vision Systems

The CCD vision system is the most common type of image sensor. With a CCD, light is captured with individual photodiode sensors. The photons that strike the sensor are converted to an equal number of electrons stored at individual sensor positions. Those electrons are then read electronically and stepped off of the charge transfer register. Once off of the CCD array, they are converted to their relative digital value.

The other type of vision sensor in digital cameras is the CMOS sensor. Both CMOS and CCD sensors are constructed from silicon. They have similar light sensitivity over the visible and near-IR spectrum. At the most basic level, both convert incident light into electronic charge by the same photo-conversion process.

CCD sensors create high-quality, low-noise images, whereas CMOS sensors are more susceptible to noise. However, during the last few years, these limitations have been sufficiently overcome as CMOS sensors do not suffer from the decrease in the signal-to-noise ratio as resolution increases. Moreover, the light sensitivity of a CMOS chip is lower and the CMOS sensors consume little power, while CCDs on the other hand consume much more power. Cost is similar at the chip level. CMOS cameras may require fewer components and less power, but they may also require post-processing circuits to compensate for the lower image quality.

IR Vision Systems

Visible light ranges from 400 nm (violet/blue) to 700 nm (red), wavelengths above 700 nm and up to about 30 μm ($\approx 30,000$ nm) are known as infrared. Infrared wavelength region is divided into near-infrared NIR (0.7–1.2 μm), mid-wave infrared (MWIR) (3–5 μm), and long-wave infrared (LWIR) (8–12 μm). Sensors working in the NIR radiation region are based on the detection of NIR photons reflected from objects, whereas the MWIR and LWIR detectors detect thermal photons which are emitted by an object originating from its heat.

Stereovision Systems

Stereovision is a technique to obtain distance information from images, which can be used in automotive safety systems. From corresponding pixels in the left and right image and the parameters of the camera configuration, the distance to a viewed 3D object can be calculated. For obstacle detection the relative height of points in the stereo images is computed from their distance. Using stereo vision detection tracking of obstacle can be accomplished.

Typical Performance of Vision Systems

First of all, range and range rate cannot be measured directly by common CCD or CMOS vision systems. They provide basically angle measurements with a resolution determined by the lenses and the number of pixels (usually VGA 640 \times 480). Approximately, the single image angle accuracy equals the angle resolution, for example, of VGA in the horizontal $\approx 50^\circ/640 \approx 0.1^\circ$. The accuracy is further determined by the quality of the lenses and the applied error calibration/compensation methods.

Methods for estimating object ranges are based on stereo imaging techniques using image disparity caused by the different image perspectives or mono imaging techniques using the optical flow. However, the range accuracy is much inferior to radar and laser scanner. Due to this, many ADAS functions are difficult to be implemented only with vision systems and a fusion with radar or laser scanner is often considered. Similar to laser scanner, range rates are also not measured directly and have to be estimated by the tracking.

The object detection performance in passive vision systems depend very much on the illumination and visibility conditions. For active vision systems, the irradiance will depend on object reflectivity, aspect angles, etc.

Major advantages of vision systems are their capability to provide edges, extensions, and poses of objects and their huge potential for object classification based, for example, on shape and surface texture recognition. However, classification performance is object and scenario dependent.

Although there is a variety of vision technologies and respective sensor systems, the general vision system characteristics are summarized in  [Table 3.4](#).

2.2.4 Ultrasonic Sensors

Ultrasonic technology is based on the propagation of sound waves for extracting information about the environment. Actually, an ultrasonic sensor transmits high-frequency

■ Table 3.4

Vision sensor specification

Detection range	3–50 m
Detection angle	Horizontal 50°
	Vertical 40°

■ Table 3.5

Ultrasonic sensor specification

Detection range	0.2–1.5 m
Detection angle	Horizontal $\pm 60^\circ$
	Vertical $\pm 30^\circ$
Frequency	43.5 \pm 2 kHz

(higher frequencies than normal hearing) sound waves and evaluates the echo which is received back by the sensor. Then it calculates the time interval between sending the signal and receiving the echo to determine the distance to an object.

Ultrasonic sensors are popular because of their low cost, lightweight, low power consumption, and low computational effort compared to other ranging sensors. For the abovementioned reasons, these sensors have the best penetration rate in the automotive market compared to other environmental perception systems and can be found in all classes of vehicles. Due to their ultrashort range (few meters), they are mainly used for parking assistance applications.

In the near future, it is foreseen that ultrasonic sensor systems will be used mostly in cheaper cars, while luxury cars will be equipped with more sophisticated sensors such as radars, lidars, and cameras.

An idea about the specifications of an ultrasonic sensor is given in [Table 3.5](#) (these are indicative values and there might be alternative solutions among different suppliers).

2.3 Virtual Sensors

The term “virtual” sensor is used for an information source which is not an actual sensor, but comprises an important input for the intelligent vehicle’s applications. The most important representatives of this category are the digital map and the wireless communication which are analyzed further in the following.

2.3.1 Digital Map

A standard digital map used in automotive applications mainly contains geometric information and other relevant attributes about the road. The core geometry consists of

links and nodes connected together forming the road centerlines of the road network. Connectivity is important for enabling routing in the network. The shape of a link, if it is not a straight line, may be represented by one or more shape points which are intermediate points between the start and end nodes of the link. As it is implied above, the shape points that describe a road segment are not placed at equidistant intervals.

All the map attributes are referenced to links, nodes, and shape points. These attributes can be points of interest (POI), traffic signs, speed limits, etc., which are sufficient for routing and navigation applications. Moreover, the map can be enhanced with further attributes such as the type of road, number of lanes, lane width, and type of lane markings which are needed for more sophisticated applications.

The digital map data can be extracted and used by a vehicle when positioning information is available. Standard map positioning techniques are based on GPS technology combined also with inertial sensors such as gyroscopes and odometers in case the satellite connection with the GPS is unavailable (Lytrivis et al. 2009a).

The accuracy of standard digital maps is difficult to be measured. Nodes and shape points are represented in World Geodetic System 1984 (WGS84) (NIMA 2000) global coordinates (latitude, longitude pairs). Coordinate resolution of current digital road maps is 10 micro degrees, which roughly corresponds to 1.1 m in latitude and 0.7 m in longitude at 50° latitude.

At this point, it should be highlighted that there might be significant differences in map data among different map vendors not only with respect to the available features and their corresponding accuracy but also in the topology of the road graphs and its completeness.

For advanced vehicular applications and especially for cooperative systems, the *Local Dynamic Map* (LDM) approach is promoted (Zott et al. 2008; ETSI 2009). The LDM is actually a map database which includes four different layers of information (see ♦ Fig. 3.2). While moving from bottom to top layers in the LDM, static information are enhanced with dynamic information (Lytrivis et al. 2009b).

2.3.2 Wireless Communication

Due to some physical limitations of the perception sensors, such as their limited range and field of view, or due to other important parameters such as their degraded performance because of bad weather conditions and their significant cost, wireless communication is examined in order to complement or even substitute these sensors and enhance the awareness of the driver. There are two types of communication used in vehicular environments: vehicle to vehicle (V2V) and vehicle to infrastructure (V2I). An overview of ITS applications that are based on the exploitation of wireless technologies is depicted in ♦ Fig. 3.3.

There are many initiatives, working groups, and organizations studying the usage of wireless communication in road environments. The most important are given below:

- *Dedicated Short-Range Communications* (DSRC) (ASTM 2003) is a short- to medium-range (1,000 m) communications service that supports both public safety and private

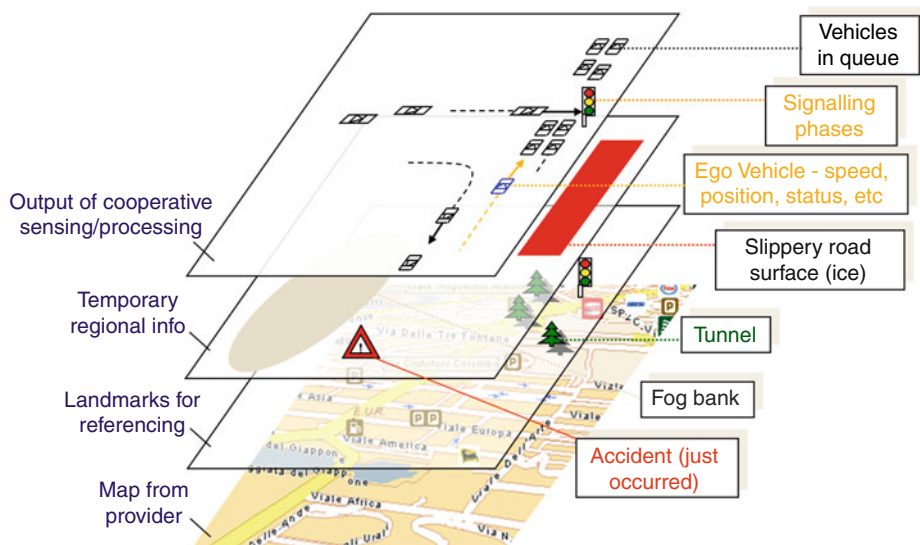


Fig. 3.2

The layered architecture of the LDM

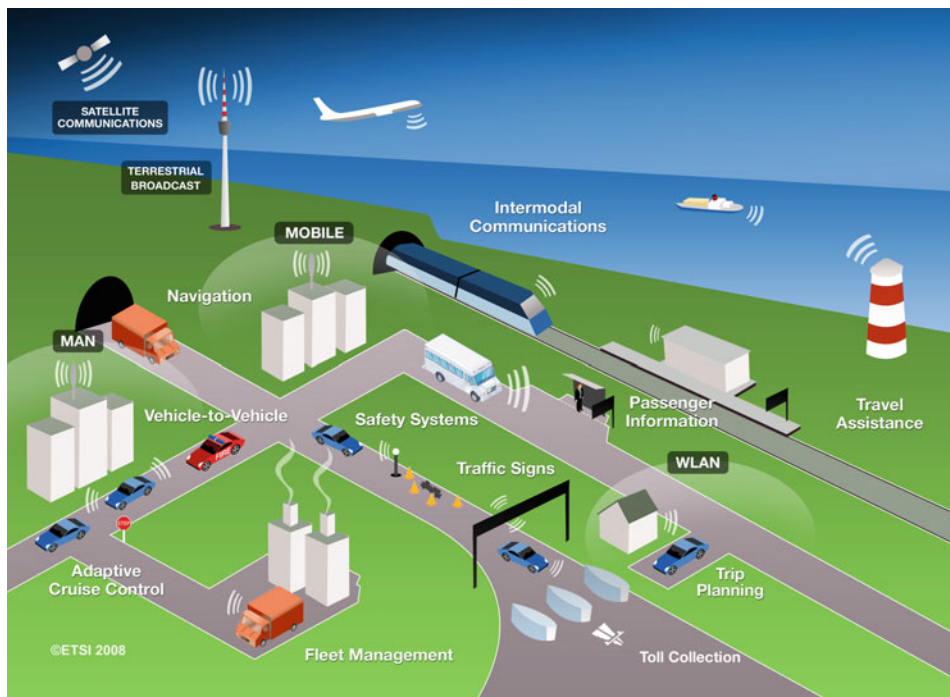


Fig. 3.3

Overview of ITS applications based on wireless communication (European Telecommunications Standards Institute)

operations in V2V and V2I communication environments by providing very high data transfer rates. It operates at 5.9 GHz and provides a spectrum of 75 MHz.

- The design of an effective communication protocol that deals with privacy, security, multichannel propagation, and management of resources is a challenging task that is currently under intensive scientific research. A dedicated working group has been assigned this specific task by IEEE, and the ongoing protocol suite is the IEEE 1609, mostly known as *Wireless Access in Vehicular Environments* (WAVE) (IEEE Standards Association 2007).
- *Continuous Air Interface Long and Medium range* (CALM) (ISO 2007; ISO TC204 WG16) provides continuous communications between a vehicle and the infrastructure using a variety of communication media, including cellular, 5 GHz, 63 GHz, and infrared links. CALM will provide a range of applications, including vehicle safety and information, as well as entertainment for driver and passengers.
- The *CAR 2 CAR Communication Consortium* (C2C-CC) (CAR 2 CAR Communication consortium; CAR 2 CAR 2007) is a nonprofit organization initiated by European vehicle manufacturers, which is open for suppliers, research organizations, and other partners. The goal of the C2C-CC is to standardize interfaces and protocols of wireless communications between vehicles and their environment in order to make vehicles of different manufacturers interoperable and also enable them to communicate with roadside units.
- *European Telecommunications Standards Institute* (ETSI) and *European Committee for Standardization* (CEN) have joined their efforts in order to provide by July 2012 common standards to be used for cooperative systems. This cooperation was based on the *European Commission Mandate M/453*.

3 Actuation

Actuators are devices that transform an input signal into motion. According to the energy source, they can be discerned into electrical motors, pneumatic or hydraulic actuators, relays, piezoelectric actuators, and thermal bimorphs. The generated motion can be linear or rotational.

Common actuators used in vehicles since long time are the fuel pump, injectors, fuel pressure regulators, idle speed actuators, spark plugs, ignition coils and controls for variable intake, cooling fan, and A/C compressor. Lately, actuators are used within the framework of more advanced systems intervening to the main vehicle subsystems, for example, within the framework of applications like ABS, electronic stability control, ACC, assisted braking, assisted steering, motor management, and chassis stabilization.

Apart from that, since the evolution of the intelligent transportation systems, actuators are being used for more sophisticated applications, that is, to actively brake the vehicle and undertake steering control in case of emergency, or even to drive the vehicle in a fully automatic way. The highly automated driving is one of the long-term visions for intelligent transport which is expected to enhance driver safety, since it is estimated that 97% of all accidents are due to human error.

3.1 Categories of Actuators According to Energy Source

3.1.1 Mechanical Actuators

Mechanical actuators convert one type of motion into another, for example, conversion of rotational into linear motion and vice versa. They are mainly used in cooperation with other actuators. An example of machine screw actuator is given in ► Fig. 3.4.

3.1.2 Electrical Actuators

Various types of motors are used as actuators in vehicles, the most common being the electronically commutated (EC) DC motors. The rotational motion produced can be used as is or converted to linear using gears or other elements.

The general requirements from motors for in-vehicle use are that they must be quiet, and resistant to vibrations, to shock, to temperature, and to chemical agents. They should



■ Fig. 3.4

Machine screw actuator by Duff-Norton (www.duffnorton.com)

also be free from electromagnetic interference with other onboard systems. The modern EC DC motors offer a very high reliability and there is no need for preset positions. Their high power density in minimum space allows their positioning in almost any location in the vehicle. Therefore, they can fit almost any vehicle design.

External rotor motors are more adequate for fans. Internal rotor motors allow the quick realization of commands due to their lower moment of inertia, thus they can be used in a wide range of applications for booster and auxiliary generating sets, for example, for controlling steering. Requirements from a motor used for steering assistance could be speed between 0 and 6,000 rpm, very low idle click point, and high uniform torque.

► *Figure 3.5* shows a smart coolant pump by Continental. It is an EC motor with integrated control electronics, which allows precise engine temperature control through variation of the coolant volume flow.

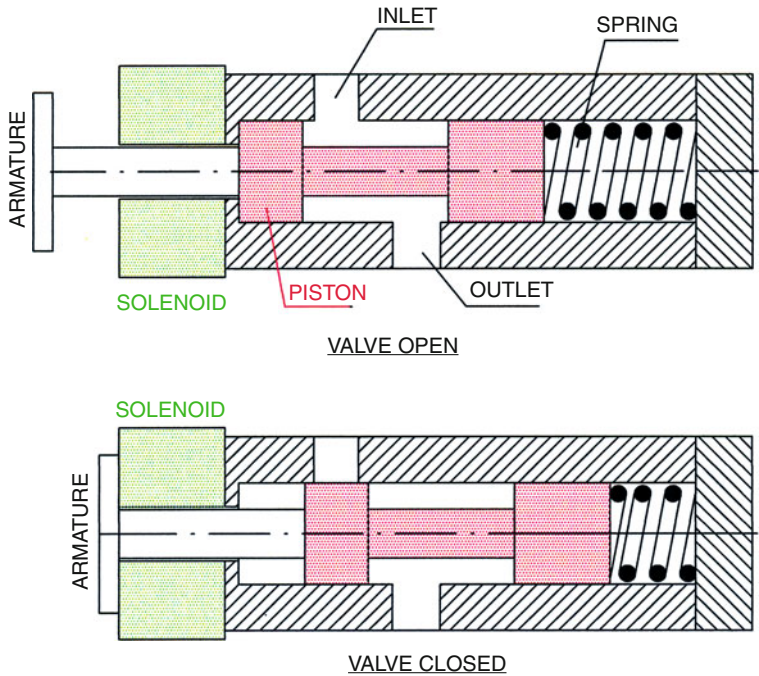
Another type of actuator commonly used in vehicles is the solenoid valves, shown in ► *Fig. 3.6*. These are pneumatic or hydraulic valves controlled by an electric current passing through a solenoid coil (a coil in the form of a helix). The solenoid is an electromechanical actuator converting electricity into mechanical movement of the pin of the valve. In some cases, the solenoid acts directly on the main valve, in others a pilot solenoid valve acts on the main valve. The latter are called pilot solenoid valves and require less power but are slower.

Another element that is commonly used as actuator in vehicles is the stepper motor, a device that converts electrical pulses into discrete movements. A drawing of a stepper motor is shown in ► *Fig. 3.7*. This is an electric motor that rotates stepwise with high precision, each step has an accuracy of 3–5% and the error is not cumulative per step. Electromagnets are arranged on the stator around a multi-toothed rotor. When the first pair of electromagnets is powered, the stator rotates one “step” until its teeth are aligned

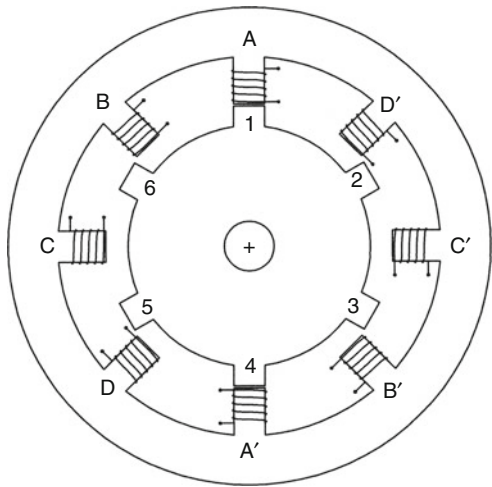


■ **Fig. 3.5**

Smart coolant pump by Continental (www.conti-online.com)



■ Fig. 3.6
A solenoid valve

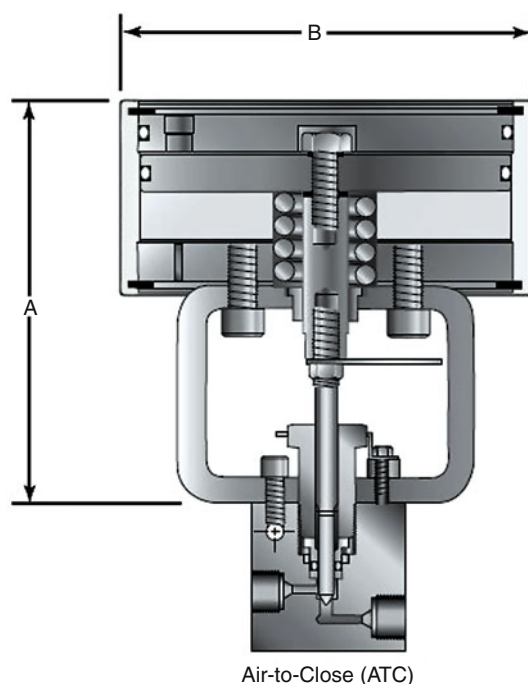


■ Fig. 3.7
A stepper motor

to the energized electromagnet, being slightly offset from the next pair of electromagnets. Then, the first pair is powered off and the second pair is powered on, thus the rotor rotates another “step” until its teeth are aligned with the second pair of electromagnets, and so on. The number of steps of movement is defined by the number of pulses. The motor speed is determined by the time delay among its electric pulse. The advantages of the stepper motor are its precise positioning and repeatability of movement, it can maintain full torque at standstill and its speed is proportional to the frequency of the input pulses.

3.1.3 Pneumatic and Hydraulic Actuators

Pneumatic and hydraulic actuators convert energy of compressed air or pressurized fluid into rotary or linear motion. They can be rod cylinders, rotary actuators, and rodless actuators with magnetic or mechanical linkage or grippers. They consist of a piston moving within a cylinder due to the pressure by the air or fluid entering the cylinder through valves. An example is shown in ► Fig. 3.8.



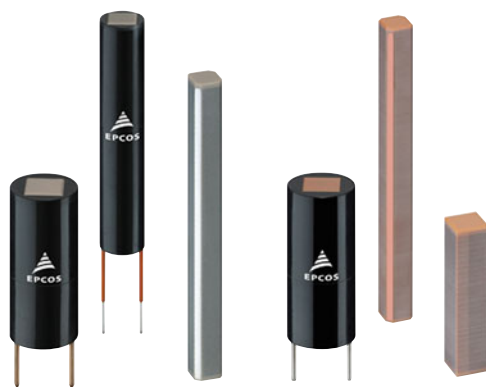
■ Fig. 3.8

Pneumatic valve actuator by Autoclave Engineers Fluid Components (<http://autoclave.thomasnet.com>)

3.1.4 Piezoelectric Actuators

The principle of operation of piezoelectric actuators is the reverse piezoelectric effect, that is, to generate a mechanical motion as a result of an applied electrical field. Such actuators are shown in ► Figs. 3.9 and ► 3.10. Actuators of this kind used in vehicles consist of multilayer ceramics, each layers being less than 100 μm in width. Amplified piezoelectric actuators can reach strokes in the order of millimeters.

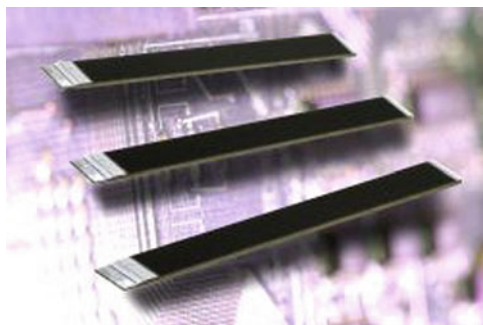
Within vehicles, such applications are mainly used for high-precision fuel injection systems replacing solenoid valves to achieve improved engine performance and reduction of fuel consumption and emissions. They can also be used for piezoelectric motors, which are considered more precise than stepper motors, and for the active control of vibrations.



■ Fig. 3.9
Multilayer piezo actuators by EPCOS (www.epcos.com)



■ Fig. 3.10
Injector driven by a piezo actuator by EPCOS (www.epcos.com)



■ Fig. 3.11
Piezo bimorphs by Morgan Technical Ceramics ElectroCeramics
(www.morganelectroceramics.com)

3.1.5 Thermal Bimorphs

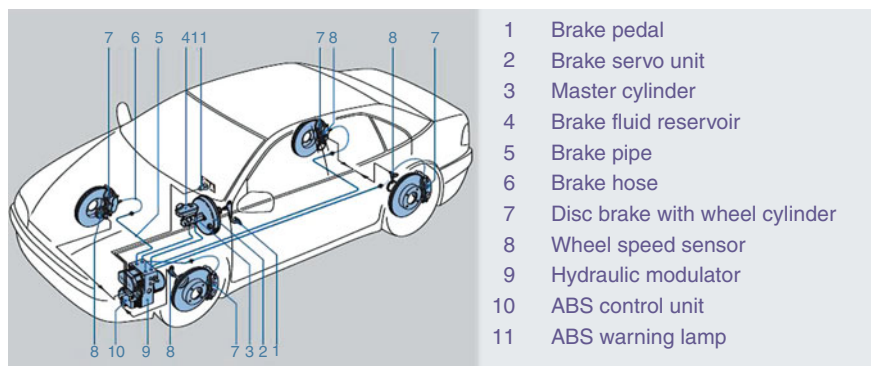
Thermal bimorphs convert electrical energy to mechanical via a thermal route. They consist of a hot and cold region, that is, a wide and a narrow limb of the same material, connected together as shown below. Electricity passing from the two regions causes a much greater temperature rise in the narrow limb, thus causing it to move.

As actuators within vehicles, they have been used as mirror position actuators, seat adjustment motors, injection valve actuators, injection pressure motors, combustion pressure motors, and headlamp position actuators for ABS systems. Example products are shown in ► Fig. 3.11.

3.2 ABS, ESC and ACC Systems

The antilock braking system (ABS) has been introduced in the market since the 1970s and its aim is to prevent wheel skidding during braking. It consists of an electronic control unit (ECU), speed sensors on each of the vehicles wheels, and hydraulic brake valves. If a slow rotating wheel is detected by the ECU analyzing the sensors signals, which signifies a wheel lock, the ECU reduces the hydraulic pressure at the corresponding wheel by activating the corresponding valve. If a fast rotating wheel is detected, then the ECU increases brake pressure to that wheel. The actuators in such systems are mainly hydraulic solenoid valves for each brake circuit, and the whole actuators unit is called the ABS modulator. Some systems have two on-off solenoid valves for each brake circuit, while others have one single valve that can operate in more than one position. Other systems, like the Delco VI ABS, use small electric motors instead of solenoids to move the valve pistons.

The major OEM suppliers of ABS are Bendix, Bosch, Delco, Continental Teves, Kelsey-Hayes, Nippondenso, Sumitomo, and Toyota. The Bosch system is shown in ► Fig. 3.12.



■ Fig. 3.12

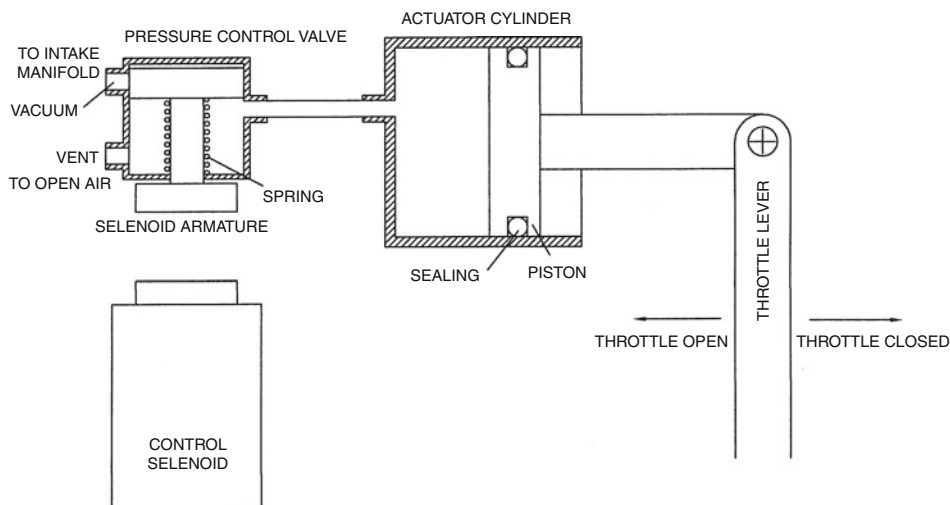
Braking system with ABS from Bosch (www.bosch.com.au)

The same or similar components installed in the vehicle for the ABS are used for other systems, which have already entered the market too. Such a system is the traction control system that detects and regulates the traction in each wheel while accelerating. Another system that controls the front-to-rear brake bias is generally called the electronic stability control (ESC) system. Such systems require two additional sensors, a steering wheel angle sensor and a gyroscopic sensor. The ESC system compares the output of the gyroscopic sensor and that of the steering wheel sensor. In case of discrepancy between the two outputs, for example, in case of skidding during emergency evasive swerves, in cases of understeer or oversteer while turning on slippery roads or in cases of aquaplaning, the ESC regulates braking pressure to the required wheels, so that the vehicle direction is in accordance to the driver's intention. On the contrary to ABS, an ESC may need to increase brake pressure more than what is provided by the master cylinder, thus an active vacuum brake booster may be used in addition to the hydraulic pump.

Another system that is already available in several models, at least as additional equipment, is the cruise control system. This system controls the vehicle speed and maintains it at the value set by the driver by acting on the vehicle throttle. More advanced systems in the same category are the autonomous or adaptive cruise control (ACC). These systems use a radar or other sensor to detect a slower moving lead vehicle and decelerate the vehicle when the relative distance becomes too short. When the traffic situation allows it, the ACC accelerates the vehicle again to the desired speed. ACC is already available for a lot of models, and it is considered the cornerstone for the future intelligent vehicles. The ACC controller regulates both the vehicle throttle and the brakes.

The throttle actuator can be operated using the manifold vacuum through a solenoid valve or with a stepper motor. The functionality of this is shown in ● Fig. 3.13.

In the first case, the amount of vacuum provided is controlled by a solenoid valve and the vacuum moves a piston that is mechanically connected to the throttle. When the solenoid is activated, the pressure control valve is pulled down and opens the path to the intake manifold pressure. In the case of ● Fig. 3.13, the lower pressure causes the



■ Fig. 3.13
Vacuum-operated throttle actuator

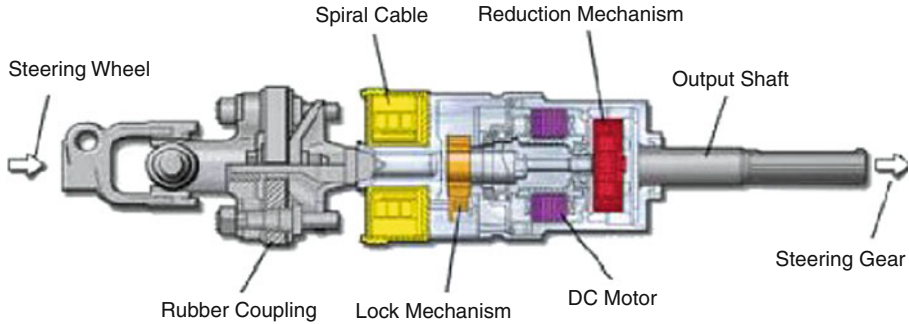
actuator piston to move to the left and the throttle opens. The extent of the throttle opening is regulated by the average pressure maintained in the chamber, which is in turn regulated by quickly switching the pressure control valve between the outside air port and the manifold pressure port.

3.3 Assisted Steering and Steer-by-Wire Systems

A lot of vehicles models are nowadays equipped with electrically assisted power steering systems, which have lately replaced the hydraulic-assisted steering systems. Such a system consists of a torque sensor in the steering column and an electric actuator supplying the adequate steering support. The existence of such components on vehicles has already opened the field for new, more advanced systems in the area.

Such systems are those that adjust the steering ratio, which is the turn of the vehicle wheels per turn of the steering wheel according to vehicle speed, which have started to enter the market. In such systems, there is no direct mechanical linkage between the steering wheel and the vehicle wheels, this is the so-called by-wire approach. Although it is technically possible to design by-wire systems as safe as mechanical systems, the by-wire systems would require significant modifications in the automotive area, this being the reason why such systems have not yet widely entered the market.

In a steer-by-wire system, the steering axle actuator can either follow driver commands, under normal driving, or follow commands from the automated system. In such systems, there should be a steering feedback to the driver via the steering wheel. The system should be able to override driver's command, if required, and for safety reasons there should be several fail-safe positions in case of system failure.



■ Fig. 3.14
BMW active steering system (www.usautoparts.net)

VW has presented a system that can fully undertake steering while parking, using the electric power steering as actuator.

The BMW active steering system, shown in ● Fig. 3.14, involves a planetary gear set integrated into the steering column.

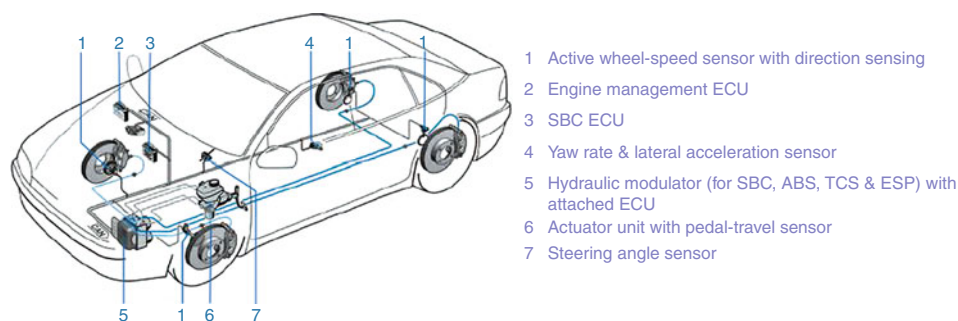
When the sensors detect a turning of the steering wheel, the system analyzes the data and sends the adequate command to an electric motor and linkage, which turn the front wheels appropriately. The actuators used in this system are a DC motor controlled by the control unit of the system and a planetary gear set between the steering rack and the steering column, which is driven by the DC motor.

At low speeds, the system makes the front wheels turn more with small angles of the steering wheel, thus facilitating parking and other maneuvers. At high speeds of around 120–140 km/h, the system reduces the change in the steering angle for every turn of the steering wheel, thus the vehicle responsiveness at high speeds is reduced and vehicle stability is improved. If skidding or sliding is detected by the yaw rate sensors, the system can adjust the steering axle of the front wheels in order to stabilize the vehicle.

3.4 Brake-by-Wire Systems

Brake-by-wire systems replace the mechanical components of a braking system with electronic sensors and actuators. Such systems are still under development and have not yet fully entered the market, mainly due to the safety-critical nature of the braking subsystem. Brake-by-wire systems have been introduced on some models by Daimler, under the name sensotronic brake control (SBC), and by Toyota, under the name electronically controlled brake.

The SBC developed by Bosch, shown in ● Fig. 3.15, substitutes the mechanical link between the brake pedal and the wheels cylinders, the brake servo unit and the brake force modulation, with an electrical link that actuates the hydraulic brake calipers. The system uses the input from a brake pedal position sensor, wheel speed sensors, a steering angle



■ Fig. 3.15

SBC components of Bosch in a car (www.bosch.com.au)

sensor, yaw rate, and lateral acceleration sensors to determine the optimum brake pressure for each wheel. The hydraulic modulator controls the brake pressure in each wheel through a high pressure pump and solenoid valves.

The electronically controlled brake of Toyota was first introduced in 2001 and is currently installed in several of its models. The brake actuator of such a system includes several solenoid valves, for the master cylinder cut, the pressure application, and the pressure reduction.

An essential safety task of the main controller of a brake-by-wire system is to monitor and control the position and speed of the brake actuators. In latest designs, this is done through the use of resolvers. In case of failure, the hydraulic system should undertake braking, so that the vehicle is always controllable.

In the future, it is expected that electronic braking systems will operate electromechanically rather than employing hydraulics, using electric motors as actuators for the brake pads.

3.5 Highly Automated Vehicles

The vision of fully automated vehicles is emerging in the last years, as a means to minimize accidents caused by driver error, which are estimated to 97% of all accidents. Still, the safety and legal problems associated with taking the driving responsibility from the human driver are a major hindrance in the market introduction of relevant systems.

Indeed, the implementation of systems that move the control from the driver to the vehicle may have an impact on driving behavior and on safety. One of the critical issues to be considered when designing such systems is the amount of control left to the driver, that is, if the driver will be able to override the automatic system, and in which case, another issue is the driver's confidence level to the system. In case of mistrust to the system due to system performance, drivers may be reluctant to handover the vehicle control. In case of driver's overtrust to the system, drivers may become overdependent to it and their own skills may wear out due to loss of familiarity with driving tasks, while being out of the loop.


On the other hand, systems that intervene in the various control tasks of driving raise numerous legal issues relevant to the liability in case of accident except of the driver of all involved stakeholders, that is, of the system and vehicle manufacturers. This is the reason why according to the Vienna Convention on Road Traffic intervening systems are only permissible, when they can be overridable by the driver, as in case of a non-overridable system the legal implications cannot be predicted at the moment.

The above have to be carefully considered when designing such systems. Below, we present some vehicles already demonstrated in real conditions and an effort to build components for the highly automated vehicle of the future.

3.5.1 Autonomous Vehicles Demonstrated in the Past

Several autonomous vehicles prototypes have been demonstrated in the past. The VITA II prototype was first demonstrated by Daimler in 1994. This vehicle could drive autonomously in highways and perform lane changes. In 1995, Daimler presented the OTTO truck that could follow a specific leader at a certain distance.

In 1997, automated vehicles for passenger transport were put at operation at the Schiphol airport.

The latest autonomous vehicles prototypes have been developed within the CyberCars initiatives. Such vehicles are designed as fully autonomous vehicles to operate at low speeds for dedicated environments. These prototypes are based on platooning, namely, close following a lead vehicle. Precise measurements of distance and relative speed to the lead vehicle are done with radar, lidar, or vision systems. Lateral positioning of the vehicle is achieved either through monitoring of the infrastructure, that is of the lane markings, or from the lead vehicle. One prototype from the CityMobil project is shown in  Fig. 3.16.

3.5.2 Advanced Automated Systems Under Development

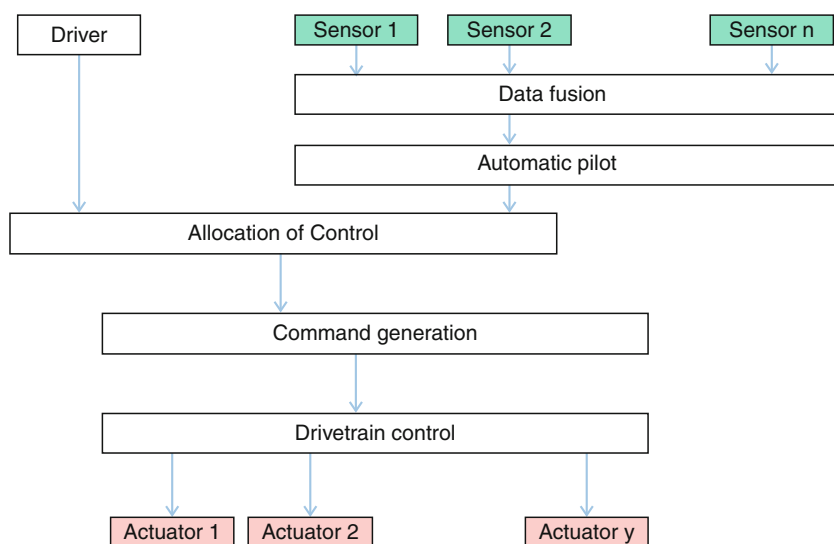
Current research activities, like the HAVEit project (www.haveit-eu.org), are focusing on the development of components and modules necessary for a fully automated vehicle. Areas that need to be researched include the task repartition between the driver and the system, so that the driver can be in the loop when required, being able to properly react in a critical situation. For this, a progressive approach to transfer the control from the system to the driver may be employed. The architecture to be followed for such future prototypes should include advanced redundancy management, so as to guarantee system availability and reliability. Applications of interest are the automated insertion into the traffic flow, the automated queue assistance, and temporary autopilot and the energy optimizing copilot.

In order to cope with the issues related to the locus of control, the architecture to be selected should include a dedicated module, which decides whether to execute or override



■ Fig. 3.16

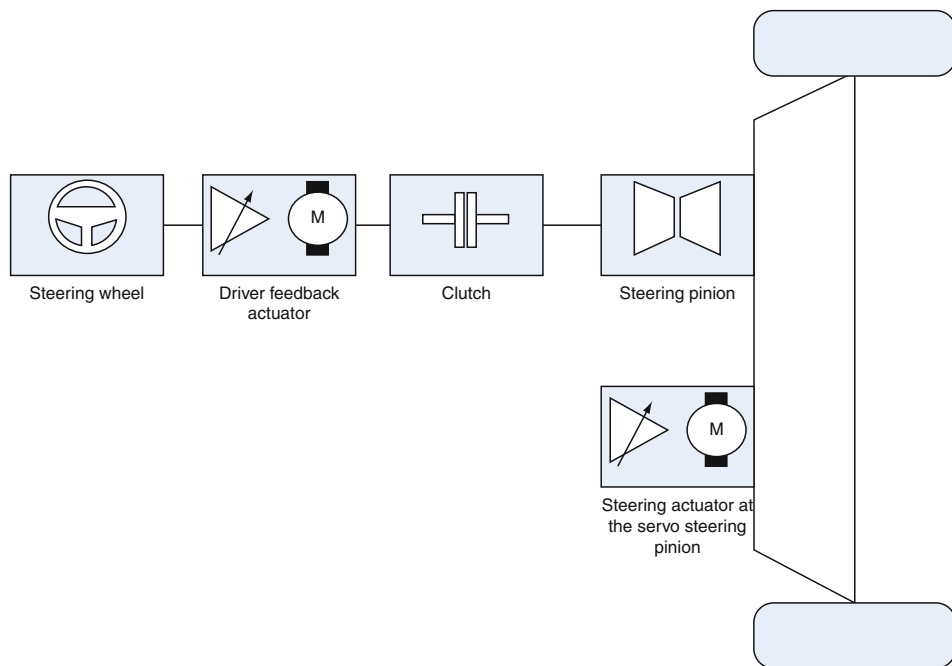
Prototype from the CityMobil project (www.citymobil-project.eu)



■ Fig. 3.17

General overview of architecture for a fully autonomous vehicle

the driver's commands, as shown in Fig. 3.17. Then, the Command Generation module will calculate the required longitudinal and lateral movement of the vehicle, and the Drivetrain Controller will decide, based on a computerized vehicle model, which actuator to activate and how, namely, the engine management, the braking or the steering system.



■ Fig. 3.18

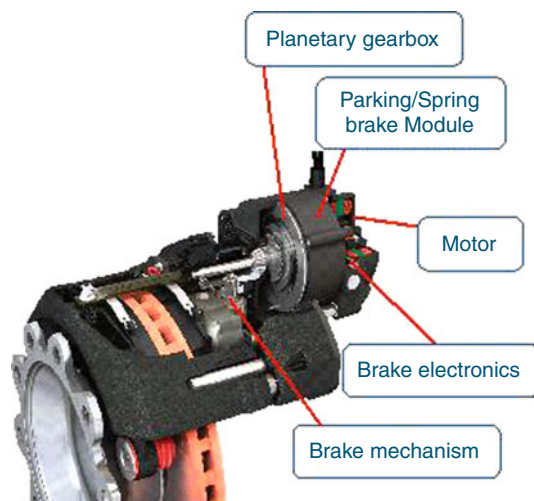
Configuration of the steer-by-wire system within HAVEit project (Jakobsson et al. 2009)

The steer-by-wire system designed for a fully automated vehicle may be configured as shown in ► Fig. 3.18.

An electromagnetic clutch may be installed between the steering wheel and steering pinion to separate steering wheel and steering axle. When the clutch remains open, steering is performed via the vehicle's servo steering actuator equipped with an adequate software. The clutch is kept open by an electromagnet. In case of power failure, the clutch couples automatically through a spring or permanent magnet, for safety reasons, so that the vehicle remains controllable through the steering wheel even in this high unlikely case.

A driver feedback actuator can be employed to generate haptic feedback to the driver, giving a steering feeling that is necessary. For this scope, an electric servo steering actuator can be used, generating torques up to 35 Nm at the steering wheel.

Regarding the brake-by-wire subsystem, it is advisable to replace the hydraulic system with the electronic wedge brake that permits to reach a level of control over each wheel brake. Such systems offer significant advantages for safety applications like ABS and ESP and they are more environmental friendly. The electromechanical brake actuator used within the HAVEit project is shown in ► Fig. 3.19. The motor used is a three-phase, brushless DC motor. The parking brake serves as an energy reserve, while the planetary gearbox transmits the motor position and generates the brake force.



■ Fig. 3.19

Electromechanical brake actuator of the HAVEit project (Nilsson et al. 2010)

This actuator converts electrical energy to force through the electric motor. Motor rotation applies clamping, while rotation in the opposite direction decreases clamping force. Motor rotational movements are transformed via a crankshaft to axial movement, pushing the pad against the brake disc. The rotational force from the disc moves the pad sideways and it starts to use the ramp resulting in an amplification of the clamping force. The applied clamping force is adjusted by the electric motor to the correct level by use of internal sensors.

4 Conclusion

The problem of road accidents and the United Nations mandate to reduce deaths by road accidents by 50% by 2020 stimulates among others the trend to develop intelligent systems to prevent accidents. On the other hand, the availability of technology at low prices makes the development of such systems more at hand.

Therefore, more and more intelligent systems are being developed, mainly for sensing the environment and enhancing the driver's awareness of the situation. Sensors already exist at affordable prices, and the effort nowadays is toward the development of more advanced sensors with higher capabilities, mainly taking advantage of cooperative techniques.

Intervening systems are not so abundant as the informatory and warning systems. This occurs due to the liability problem of who will be blamed in case of an accident and because a severe system failure may have fatal results. Fail-safe design and redundant circuits are a necessity for intervening systems, but they increase their cost and thus their possibility for market success.

In the future, we may be entering a car which will be driven by an automatic pilot. This will be for sure very comfortable for the driver and all passengers, it will reduce driver's fatigue even driver's frustration in cases of traffic jams, and it will be safer. But will it be fun?

References

- ASTM (2003) Standard specification for telecommunications and information exchange between roadside and vehicle systems-5 GHz band dedicated short range communications (DSRC) medium access control (MAC) and physical layer (PHY) specifications, September 2003. ASTM, West Conshohocken, PA
- CAR 2 CAR communication consortium (2007) C2C-CC manifesto, version 1.1, August 2007. <http://www.car-to-car.org/>
- CAR 2 CAR Communication consortium (C2C-CC). <http://www.car-to-car.org/>
- Co-operative systems for road safety "Smart vehicles on smart roads," SAFESPOT, FP6 integrated project. <http://www.safespot-eu.org/>
- ETSI TR 102 638 v1.1.1, Intelligent Transport Systems (ITS) (2009) Vehicular communications; Basic set of applications; Definitions, June 2009. ETSI, Sophia Antipolis, France
- European Commission Mandate M/453. http://ec.europa.eu/information_society/activities/esafety/doc/2009/mandate_en.pdf
- European Committee for Standardization (CEN). <http://www.cen.eu/cen/pages/default.aspx>
- European Telecommunications Standards Institute (ETSI), Intelligent Transport Systems (ITS). <http://www.etsi.org/website/Technologies/IntelligentTransportSystems.aspx>
- European Telecommunications Standards Institute (ETSI). <http://www.etsi.org/WebSite/homepage.aspx>
- <http://autoclave.thomasnet.com/item/pneumatic-valve-actuators/pneumatic-valve-actuators-2/item-3669>
- http://www.bosch.com.au/content/language1/downloads/brakepads_2005.pdf
- <http://www.citymobil-project.eu/site/en/SP1%20Rome.php>
- <http://www.duffnorton.com/products/specs.aspx?catid=1922>
- <http://www.morganelectroceramics.com/products/piezoelectric/piezo-bimorphs/>
- <http://www.usautoparts.net/bmw/technology/afs.htm>
- http://www.conti-online.com/generator/www/de/en/continental/automotive/themes/passenger_cars/powertrain/sensors_actuators/sensors_actuators_en,tabNr=2.html
- <http://www.epcos.com/web/generator/Web/Sections/Components/Page,locale=en,r=263288,a=263406.html>
- <http://www.epcos.com/web/generator/Web/Sections/ProductCatalog/CeramicComponents/MultilayerPiezoActuators/Page,locale=en.html>
- IEEE Standards Association (2007) IEEE P1609.1 – Standard for wireless access in vehicular environments (WAVE) – Resource manager, IEEE P1609.2 – Standard for wireless access in vehicular environments (WAVE) – Security services for applications and management messages, IEEE P1609.3 – Standard for wireless access in vehicular environments (WAVE) – Networking services, IEEE P1609.4 – Standard for wireless access in vehicular environments (WAVE) – Multi-channel operations, adopted for trial-use in 2007, IEEE Operations Center, 445 Hoes Lane, Piscataway, NJ
- International Organization for Standardization (2007) Intelligent transport system-continuous air interface long and medium (CALM) – Medium service access point. Draft international standard ISO/DIS 21218, ISO, Geneva, Switzerland
- ISO TC204 WG16. <http://www.isotc204.com>
- Jakobsson E, Beutner A, Pettersson S et al (2009) HAVEit project deliverable 12.1 "Architecture," February 2009. www.haveit-eu.org
- Low-cost miniature laserscanner for environment perception, MiniFaros, FP7 small or medium-scale focused research project (STREP). <http://www.minifaros.eu/>
- Lytrivis P, Thomaidis G, Amditis A (2009a) Sensor data fusion in automotive applications. In: Nada Milisavljevic Ir (ed) Sensor and data fusion. I-Tech Education and Publishing KG, Vienna, Austria, pp 133–150. ISBN: 978-3-902613-52-3
- Lytrivis P, Vafeiadis G, Bimpas M, Amditis A, Zott C (2009) Cooperative situation refinement for

- vehicular safety applications: the SAFESPOT approach. In: ITS World congress 2009, Sweden, 21–25 Sept 2009, Stockholm, Sweden
- Marek J et al (2003) Sensors for automotive applications. (Sensors Applications Vol 4). Wiley, Weinheim
- Nilsson A, Nilsson P, Seglő F (2010) HAVEit project deliverable 23.1 “Brake-by-Wire for challenge 4.2,” January 2010. www.haveit-eu.org
- NIMA (2000) Department of Defence World Geodetic System 1984 – Its definition and relationships with local geodetic systems. Report TR8350.2, 3rd edn. National Imagery and Mapping Agency, Bethesda, MD
- Preventive and active safety applications, PREVENT, FP6 integrated project. <http://www.prevent-ip.org/>
- Zott C, Yuen SY, Brown C, Bartels C, Papp Z, Netten B (2008) SAFESPOT local dynamic maps – context-dependent view generation of a platform’s state & environment. In: 15th World Congress on ITS 2008, New York, November 2008

4 Situational Awareness in Intelligent Vehicles

Zoltán Papp

TNO Technical Sciences, The Hague, Netherlands

1	<i>Introduction</i>	62
2	<i>The Big Picture: Situational Awareness for Intelligent Vehicles</i>	63
3	<i>Situational Awareness and Communication</i>	67
4	<i>World Modeling, Representation</i>	69
5	<i>Situational Awareness in Control</i>	74
6	<i>Conclusions</i>	78

Abstract: Cooperative intelligent vehicle systems constitute a promising way to improving traffic throughput, safety and comfort, and thus are the focus of intensive research and development. The vehicles implement more and more complex onboard functionalities, which interact with each other and with their surroundings, including other vehicles and roadside information infrastructure. In order for a functionality “do the right thing” it should have a sufficiently complete and certain interpretation of the surrounding world (i.e., relevant part of the road infrastructure, the surrounding vehicles, the ego-vehicle itself, etc.). Due to the ever-existing limitations of the sensing, the complexity of the data interpretation and the inherent uncertainty of the world “out there”, creating this representation poses major challenges and has far reaching consequences concerning how onboard functionalities should be built. The situational awareness term covers an overarching research field, which addresses this understanding process from different angles. It attempts the conceptualization of the problem domain, it relates sensory data processing and data fusion with the understanding process, and it investigates the role of humans in the related processes and even gives architectural guidelines for system design.

First a brief overview is given about the established models for situational awareness emphasizing the specialties of the intelligent vehicle systems. Then the representation problem is covered in details because the representation has strong influence both on the sensing, data interpretation, control and architectural aspect. Finally the control and architectural aspects are covered addressing the design for dependability.

1 Introduction

The ever-increasing sophistication of onboard functionalities helps to cope with the ever-increasing challenges of improving traffic throughput, safety, and comfort. Establishing cooperation among the players and thus creating intelligent vehicle systems is the next frontier of developments transforming the way how drivers are involved in the driving tasks and how traffic is managed. Thus intelligent vehicle systems are the focus of intensive research and development (Varaiya 1993; Lygeros et al. 1998; Horowitz and Varaiya 2000; Halle et al. 2004; Li et al. 2005a, b; Michaud et al. 2006). The state-of-the-art intelligent vehicle applications usually can be described as a collection of highly autonomous, complex dynamical systems interacting on various functional levels. The set of functionalities implemented onboard covers a wide spectrum in complexity and sophistication: the simple ones implement “single minded” control functions (e.g., cruise control); the more sophisticated ones can automate complex driver tasks (e.g., cooperative adaptive cruise control and parking assistant). These functionalities manifest different levels of autonomy: certain functionalities run without human interaction (e.g., ABS), others act as an advisory function leaving the action to the driver (e.g., lane departure warning).

In order for a functionality “do the right thing” it should have a sufficiently complete and certain interpretation of the surrounding world (i.e., relevant part of the road infrastructure, the surrounding vehicles, the ego-vehicle itself, etc.).

This more or less obvious statement has far-reaching consequences. Though typically an intelligent vehicle is richly instrumented with various sensors (e.g., different types of radars, proximity sensors, accelerometers, and gyroscopes) and ever-powerful computing platforms, the “understanding” process poses extreme challenges for researchers and developers. This is due to the ever-existing limitations of the sensing, the complexity of the data interpretation and the inherent uncertainty of the world “out there”. The situational awareness term covers an overarching research field, which addresses this understanding process from different angles. It attempts the conceptualization of the problem domain, it relates sensory data processing and data fusion with the understanding process, and it investigates the role of human in the related processes and even gives architectural guidelines for system design. In the following the aspects of situational awareness and – because in the application domain considered the awareness is not the “result” but only an intermediate stage to reach our main goal, which is management of the situation (automatic or not) – the related aspects of control are overviewed from the intelligent transportation system’s point of view.

Though situational awareness problems surface in each component comprising the intelligent transportation system, the article considers the situational awareness problem from the intelligent vehicle’s point of view. First a brief overview is given about the established models for situational awareness (mainly originated in the defense application domain). The specialties of the intelligent vehicle systems are summarized next emphasizing the relaxing and challenging aspects. As it will be shown the situational awareness has various “ingredients,” such as sensing, sensory data interpretation, data fusion, communication, representation, etc. This chapter applies a representation centered view. The representation is a key issue in situational awareness and has strong influence on the sensing, data interpretation, control, and architectural aspects. Consequently it determines system level properties as robustness, efficiency, maintainability, etc. The other “ingredients” are covered elsewhere comprehensively (including this volume). After covering the representation aspect the control and architectural aspects are considered: the main challenges are overviewed the “primary functions” face. Here the “primary functions” refers to the functionalities the onboard system should deliver to the user (i.e., the driver). Everything, which comes before this in the processing chain (i.e., including all functionalities involved in creating situational awareness) is merely a necessity to serve the needs of the primary functions. The section overviews how the primary functions should be designed in order to assure dependability. The chapter concludes with summarizing situational awareness-related design guidelines for cooperative vehicle systems and indicates the most immediate research challenges.

2 The Big Picture: Situational Awareness for Intelligent Vehicles

Future vehicles host more and more advanced driver support functionalities to enhance traffic throughput, safety, and comfort. These functionalities cover a wide spectrum in

complexity, safety criticality, time criticality, autonomy, etc. As the functionalities become more complex, autonomous, and control centered (i.e., taking over certain driver functionality for achieving partially automatic driving or implementing functionalities human drivers are typically not really good at, such as pre-crash control), the requirements for dependable operation become more stringent and represent a primary concern for system design (Lygeros et al. 1996; Horowitz and Varaiya 2000). (In the following dependability is used as an “integrating term” for robustness, fault tolerance, graceful degradation, and fail safety. When relevant the scope and degree of the dependability is given more accurately in the application context.)

The onboard functionalities considered are control functionalities: they generate control commands based on the observations. The control commands – via actuators – influence the behavior of the hosting vehicle. The control loop is closed either directly (i.e., the actuator acts on the vehicle) or via informing the driver who acts as an actuator (consequently becomes an element of the control loop). Vehicles operate in a complex, unstructured and sometimes hostile environment. Onboard sensory systems (e.g., radar, laser rangefinder, accelerometers, and wheel encoders) and communication links (car-to-car or car-to-infrastructure communication) are used to acquire information about the vehicle itself and about its relevant surroundings. The “meaning” of the sensory readings and incoming messages are “extracted” by local (i.e., onboard) data processing. Deriving the “meaning” is one of the most important aspects of dependable performance: effective and safe control commands can only be generated if the relevant part of the environment is known with high certainty. The basis of robust and safe behavior is the understanding of the circumstances the vehicles (and other relevant players of the given configuration) are in. **Figure 4.1** shows the conceptual processing scheme. The *World Model* is the interface between the sensory data interpretation and the control functionalities. (Here and in the following the “control functionality” term is used in a wide sense. The term covers continuous and discrete control, decision making, advice generation, etc. In the latter case the driver can be considered as “actuator” (with particular dynamics and built-in control capabilities).) The *World Model* is the explicit representation of the *World* as the

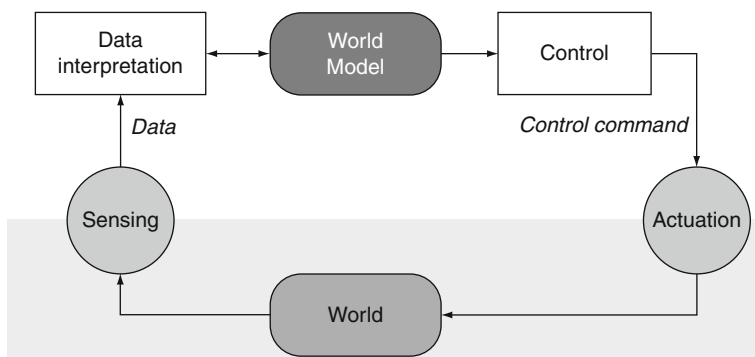


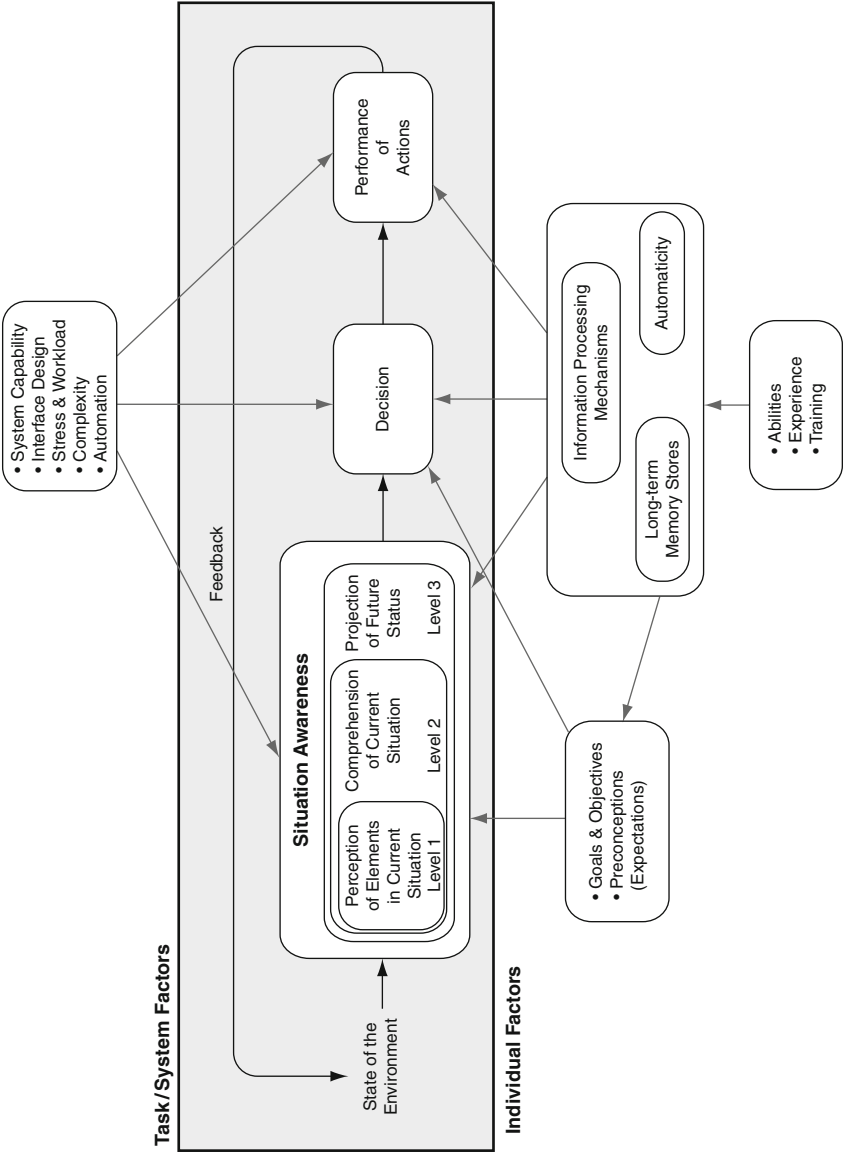
Fig. 4.1
Conceptual system decomposition

vehicle “knows” it. The sensory data interpretation side – using all available observations – builds this representation: it identifies real-world objects, determines their attributes, and stores a formal representation of these in the *World Model*. The *World Model* contains the representation only of a subset of the objects in the world. If the sensory system is adequate and the incoming observation stream is correctly interpreted, then the *World Model* properly represents the relevant part of the surrounding world (i.e., with small enough uncertainty, sufficient accuracy).

The problems of building an adequate world model are targeted by the field of situational awareness (SA). One of the most widely accepted definition of SA is given in (Endsley 1995). Here SA is defined as the *perception* of elements in the environment within a volume of time and space, the comprehension of their *meaning*, and the *projection* of their status in the near future. Note that SA is inherently limited both in space and time. In every time instant SA can be considered as a “state of knowledge”. The process, which creates SA is called *situation assessment*. It is important to realize that situational awareness and situation assessment have a dynamical relationship: the actual state of SA may have influence on the assessment process in the coming time period (e.g., it can determine what algorithms are considered applicable to improve SA) and the assessment process influences the updated SA. Early work on SA mainly considered human centered defense related systems – and it is reflected in the common terminology, case studies and examples. A commonly accepted model of (human centered) decision making – with the role of situational awareness indicated in it – was introduced by Endsley (Endsley 2000). Three stages of the situation assessment creating awareness are distinguished in this model (● Fig. 4.2).

- Level 1 – Perception: Perception is the lowest level of information gathering, refers to acquiring cues (“features”). Perception is an essential step because this determines the information content of the situation assessment process. Every following stage performs merely information extraction from this raw data set.
- Level 2 – Comprehension: On this level the integration of the information (provided by the perception) takes place. The relevance of the data (with respect to the goals and objectives) is determined and relationships discovered.
- Level 3 – Projection: Situation assessment is not only about the present. The projection (i.e., prediction) from the current situation taking into consideration the dynamics of the observed phenomena and the anticipation of future events (i.e., considering future implications) constitute the highest level of situational awareness.

In early systems the emphases were placed on the data acquisition, mainly Level 0 signal processing and presenting the result to the operator in a proper form. (Consequently the human-machine interface played a pivotal role. A significant body of the situational assessment research was carried out in this domain.) With the advance in computing and with the spread of distributed systems, ubiquitous sensing and mobile communication, the extent of the automated situation assessment grew dramatically encouraging novel developments in signal processing, data fusion, knowledge representation, etc. These developments were also motivated by application challenges: the “traditional” human-in-the-loop decision making (control) systems could not be used to further



■ Fig. 4.2
Model of situational awareness in dynamic decision making (From (Endsley 2000))

increase the performance of the attached process. This architecture resulted in information overload on the human side, thus constraining the performance. These issues profoundly surfaced first in the defense application domain. In order to facilitate development and interoperability the conceptualization of the information abstraction led to a commonly used data fusion model, the JDL model, which can be easily matched with situation assessment process. (Process-wise the assessment process at least partially manifests itself in the data fusion process because the assessment uses various sensory inputs about the observed phenomena and the sensory data are interpreted jointly resulting in deeper insight (Hall and Llinas 1997).) The JDL model also evolved in time, but the principles remained (Steinberg et al. 1999): the model defines abstraction levels for the assessment related to the problem-space complexity. In the baseline JDL models the following levels are defined:

- Level 0: estimation of the signal features and properties in sub-object level (“signal assessment”)
- Level 1: estimation of the states of physical or conceptual objects (e.g., vehicles, weather) (“object assessment”)
- Level 2: determining the relationships among objects, such as “followed-by,” “acting-on” (“relationship assessment”)
- Level 3: estimation of the impact, i.e., the consequence of the given situation on the evolution of the phenomena observed and on the user goals (“impact assessment”)

Though the concepts belonging to these levels have a stronger “engineering flavor” than in the Endsley situation assessment model, the parallel between the models are obvious. The intelligent vehicle systems and mobility application domain bear fundamental similarities to the defense domain as far as the situational awareness aspect is concerned, thus the results of the defense research are readily “portable” in the mobility domain.

Both domains are characterized by

- Relying on uncertain observation streams characterizing an uncertain world
- Applying real-time reasoning in a dynamic context to achieve (partial) understanding
- Ability to make decisions (control command calculations) in a guaranteed time window always making optimal use of the available insight into the controlled process
- Assuring dependability

(Polychronopoulos and Amditis 2006) considered JDL as a foundation and extended, dedicated the JDL model to automotive (safety) applications by attaching automotive specific terms to the JDL levels and detailing the inter- and intra-layer interaction for e-safety applications.

3 Situational Awareness and Communication

In the recent years the widespread availability of affordable and high-speed wireless communication transformed the situational awareness landscape and gave impulse to

the development of novel cooperative mobility applications (Caveney 2010). This transformation role is due to the following characteristics:

1. The communication extends the range of traditional sensing. There is no need for line of sight, optical occlusion does not mean “invisibility,” etc.
2. Information can be shared easily among vehicles and roadside infrastructure and typically has better quality (i.e., higher accuracy, less uncertainty) than the information derived from local sensors.
3. The communication is a two-way channel, i.e., beside the “passive sensor channel” (feeding in information from other entities) it can be used to emit messages – providing the basic mechanisms for interaction and coordination.

Unfortunately the communication brings in a new class of uncertainties, too. The eventual drop-out of communication and the failing communication equipment introduces new failure modes, which should be handled in the situation assessment process. It should be emphasized that these failure modes are different than that of the onboard sensory system. The traditional “onboard redundancy” cannot provide a solution here: a vehicle with failing communication can completely disappear from the “map.” As a fall-back scenario onboard sensory systems should take over observation functionalities, but in most cases the range and accuracy are severely limited – which have far-reaching effects on the higher-level functionalities the system can safely offer. (The considerations for the proper functional design of control and decision making operations are addressed later in the chapter.)

The communication enables creating distributed systems via connecting functionalities implemented on different platforms by different vendors eventually in a dynamically changing environment – thus the importance of the interoperability aspects cannot be overemphasized. Industry and government supported consortia in Europe, USA, and Japan work on identifying application classes, the related communication requirements and standards (Caveney 2010; NHTSA 2002–2005; [Car 2 car communication consortium manifesto](#)).

Intelligent transportation systems are complex spatially distributed dynamical systems with a great number of stakeholders participating in their operation. In order to assure situational awareness, information should be gathered from diverse sources such as traffic management centers, weather stations, city event calendars, individual vehicles, emergency crews, etc. Obviously, the ubiquitous communication makes these connections feasible. On the other hand the rapid deployment of sensors and sensor networks hosting wide variety of sensors with diverse capabilities and allocation makes the sensory data interpretation task a challenge. Current standards assume that the “meaning” of the data contained in a message is known implicitly. This approach leads to isolated “islands of data” and a rigid, fragile information infrastructure, which limits innovation and the introduction of new applications and services.

An emerging and promising related field is the semantic sensor web (SSW), which addresses this problem via defining a framework for providing the meaning of the observations (Sheth et al. 2008). The semantics is attached to the data content as meta-data and thus enables the correct interpretation of the data. The semantic sensor web

development is an integration, extension, and customization of the Sensor Web Enablement (SWE) (Botts et al. 2007) and W3C Semantic Web (<http://www.w3.org/standards/semanticweb/>) efforts resulting in dedicated solutions for creating semantically rich sensor networks and thus enabling situational awareness. The key elements of the SSW are ontologies developed for particular application domains. Ontology is a formal representation of a domain: it defines the concepts and their relationships used to describe application instances. SSW uses three connected ontologies for spatial, temporal, and thematic characterization of observations. Various standard organizations, trade groups, scientific associations, etc., work on dedicated SSW ontologies to cover the special needs of the targeted domain (e.g., Open Biomedical Ontologies, Geographic Markup Language Ontology, NIST Standard Intelligent Systems Ontology).

4 World Modeling, Representation

The situation assessment process results in situational awareness, i.e., all relevant objects and their relations in the relevant space-time volume are identified. The typical signal flow of the process is shown in Fig. 4.3.

Onboard sensors (S) and communication links (C_R for receiving) are the source of the observations; the static information is represented by the MAP component (i.e., road geometry, landmarks, etc.). It is important to realize that conceptually the communication is handled as sensory input, i.e., the received data sets run through the situation assessment process. The result of the assessment process is stored in the *World Model*, which serves as the sole interface to the decision making and control functionalities

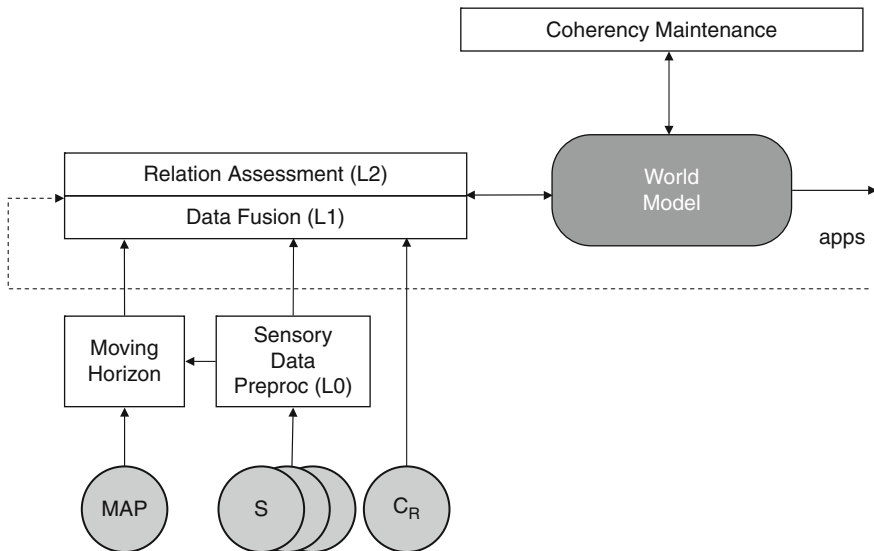


Fig. 4.3

From sensing to world model

(labeled as applications). Note that the applications may derive “extra understanding,” which should extend the *World Model* – thus these findings also run through the situation assessment as the situation assessment is responsible for considering all relevant inputs to create awareness. The processing scheme showed has a number of advantageous features:

- The *World Model* serves as an abstract interface between the sensory system and the control functionalities and thus assures certain independence between them.
- Upgrading the sensory and/or the communication subsystems typically results in a more accurate “world view” (e.g., higher accuracy, extended range). The improved “world view” can immediately improve control performance (i.e., without modifying the control functionalities) provided the control functionalities are able to handle the meta-information attached to derived world attributes (e.g., ranges and uncertainties, see details later).
- The architecture shown above inherently supports creating robust onboard systems. Achieving robust operation is not without conditions: both the sensory data interpretation and the control side should be properly constructed; upcoming sections detail this aspect of the system design.
- In cooperative vehicle systems one of the roles of the communication is to improve the “world view,” i.e., to give a better foundation for control. The architecture shown can easily be extended with “inter *World Model*” communication creating the possibility for sharing local world models among vehicles. Via the application of data fusion and consensus building algorithms a shared situational awareness can be created.

Consequently the *World Model* plays a pivotal role in this configuration. It has profound effect on the performance: what is not represented in the *World Model* cannot play a role in the higher-level functionalities, i.e., the representation used by the *World Model* should be rich enough to reflect all relevant objects and their relations. Novel developments in world modeling for intelligent vehicle systems explicitly consider the needs of situational awareness. (Schlenoff et al. 2005) makes an attempt to formalize the 4D/RCS reference model architecture (Albus 2000) developed for autonomous unmanned vehicles. By providing a standard set of domain concepts along with their attributes and interrelations the work facilitates knowledge capture and reuse, systems specification, design, and integration. (Matheus et al. 2003a, b) strictly focuses on the situational awareness related aspect of world representation and proposes an ontology for characterizing a world populated with physical objects with spatiotemporal behavior.

In the following – in order to illustrate the abstract concepts with an example – the principles of the world modeling are shown via a particular implementation developed and used in the SAFESPOT Integrated Project (<http://www.safespot-eu.org/>). The SAFESPOT project’s main goal is to improve traffic safety by applying cooperative systems. In applications relying on SAFESPOT technologies, vehicles and the road infrastructure communicate to share information gathered on board and at the roadside to enhance the drivers’ perception of the vehicle surroundings, i.e., SAFESPOT builds cooperative awareness. (The SAFESPOT demo applications did not close the control loop, i.e., the applications acted as advisory systems to the driver, who was fully in charge

of controlling the vehicle. On the other hand the technologies developed in the SAFESPOT project are also applicable in automated control configurations.) In the SAFESPOT terminology the world model is called Local Dynamic Map (LDM). LDM is part of the onboard platform of every SAFESPOT vehicle and responsible for maintaining the formal representation of the world surrounding that vehicle (Papp et al. 2008; Zott et al. 2008).

The LDM contains the object oriented representation of the world constructed by the data interpretation system based on static map data, sensory and communication inputs (● Fig. 4.3). This representation is an instantiation of an object model as defined by a custom ontology. The concept and the structure behind the object model is generic, an actual implementation of the object model depends on the application domain to be covered. ● Figure 4.4 shows the top-level segment of the class diagram describing the LDM object model. The main characteristics of the object model are as follows.

Hierarchy, inheritance: The LDM uses object oriented approach to represent the model of (the relevant regions and aspects of) the real world. The object hierarchy (specialization) and inheritance are common object oriented design concepts, which greatly reduce the complexity of the description and enhance its expressiveness at the same time. The *WorldObject* is the most generic (abstract) object. Every concrete object in LDM extends *WorldObject* in an indirect way: an object is an instance of either a *StaticObject*, *DynamicObject* or *CompoundObject* (or their further specialized variants).

Position, motion state: In order to characterize the motion state of an object a coordinate frame (coordinate axis system) is attached to the object and its 6 degrees of freedom motion state is expressed as the motion of this attached frame with respect to a reference frame. The reference frame can be either an “absolute world frame” (WGS84 standard) or the attached frame of any other world object. This latter case is used to characterize relative motions: a vehicle is seen from the viewpoint of another (used as reference). Objects can be represented in different frames simultaneously.

Geometry: The spatial properties are essential for most objects playing a role in the traffic scenario. The thorough and efficient representation of the geometry and related properties are critical because certain object relations can only be deduced and maintained via operations on spatial properties. The geometry of the objects is defined by the associated *GeometryObject* class and its subclasses. The separation of the geometry from the object was partially driven by implementation characteristics: namely, extensive libraries and database extensions are available to handle spatial features. Object geometry is described relative to the embedded frame of the object itself.

Conceptual objects: As mentioned earlier everything to be represented in the LDM should be an object even if it cannot be considered as a real-world physical object – and thus are called conceptual objects. Events, traffic, and environmental phenomena are typical examples of conceptual objects. In the LDM there are instances of the *ConceptualObject* type (or one of its subclasses). Many times *ConceptualObjects* cannot be observed directly, rather they are discovered/derived by the higher-level data interpretation functionalities of the situation assessment process (e.g., traffic congestion can be deduced and its parameters can be calculated via observing the motion state of clusters of vehicles).

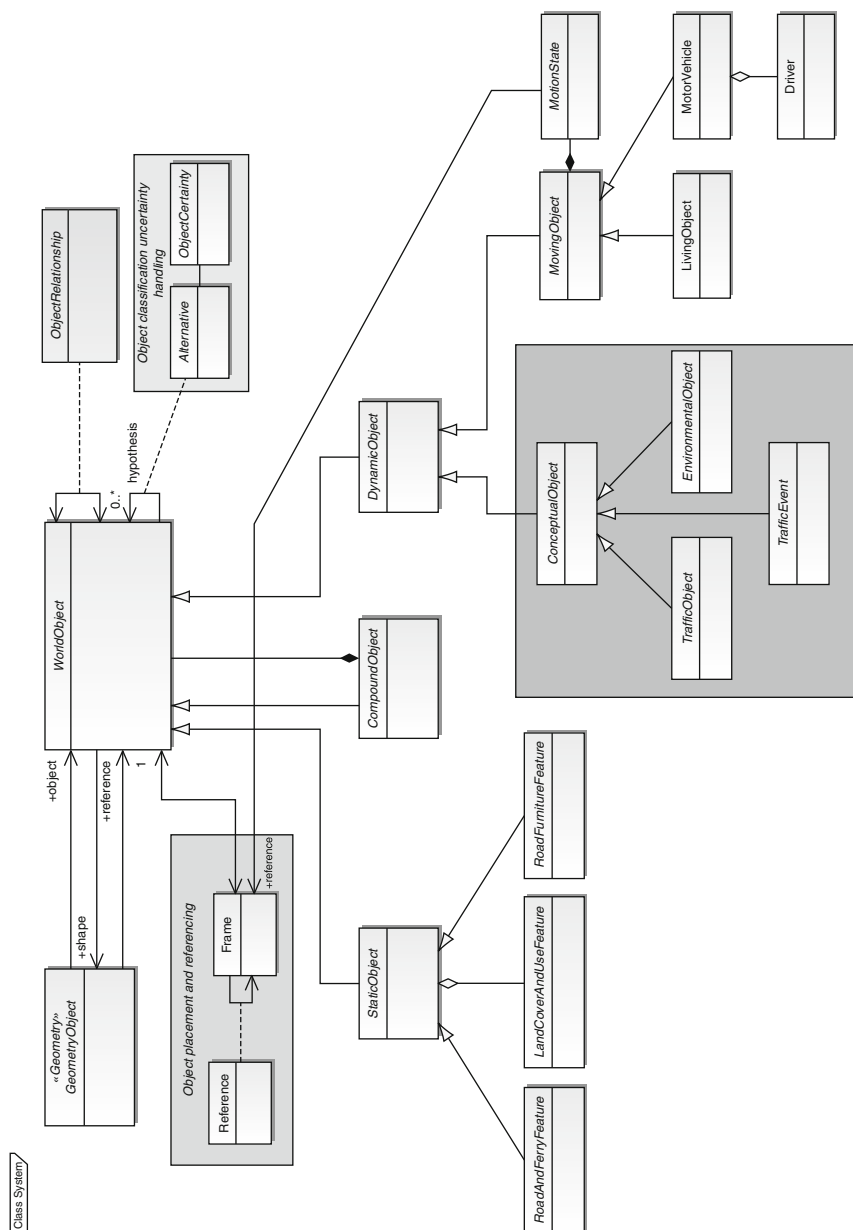


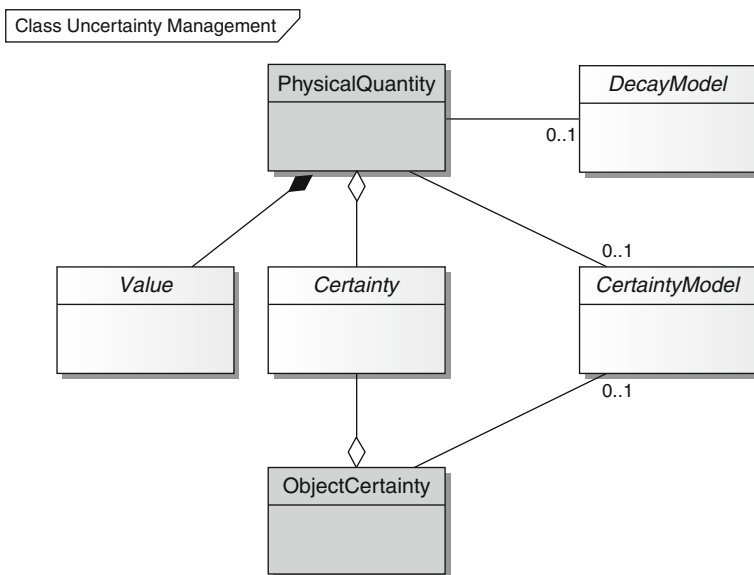
Fig. 4.4

The SAFESPOT LDM object model

Object relations: Many times not the attributes of the individual objects but the relations between the object characterize the situation. The appearance or disappearance of a particular relation (e.g., being behind a car) and the attributes of the relation (e.g., the actual headway in the previous example) constitute the relevant information. Some (typically spatial) relationships can be deduced on-demand from the LDM contents. However, deriving certain relationships on-demand can be very expensive, e.g., it would require running extensive search algorithms whenever new sensor readings are available. Relations can be “explicitly represented in the LDM in a static or dynamic way” (e.g., vehicle runs on a particular road segment, the traffic light belongs to a lane). This is advantageous if maintaining the relation is less expensive computationally than establishing it. Relationships are described via concrete subclasses of the *ObjectRelationship* type. *ObjectRelationship* can relate any number of *WorldObjects*.

Uncertainty: The LDM can represent two types of uncertainties (➤ Fig. 4.5).

- *Object uncertainty* is the result of ambiguous object classification. The available sensory input – due to its limited range and/or observation noises – may not be sufficient to determine the type of the real-world object (e.g., it cannot be decided whether the readings correspond to a single truck or a group of occluding passenger cars). In these cases, LDM stores alternative representations simultaneously with attached certainty descriptors (see ➤ Fig. 4.4). The attributes of the certainty descriptors are updated as new observations become available. Alternatives get removed from the LDM when the certainty descriptor indicates unambiguous classification is reached.



■ Fig. 4.5
Uncertainty representation

- Measurement and modeling errors result also in *attribute uncertainty*. Attribute uncertainty characterizes the “trustworthiness” of the attribute, i.e., it influences every control action and decision using the attribute. Consequently the uncertainties should be propagated through the data processing chain. Certain control functionalities cannot be executed whenever the input uncertainties are too high (see more details in the next section). Every attribute derived from sensory inputs or resulting from uncertain decisions are represented as *PhysicalQuantity* (► Fig. 4.5). The *PhysicalQuantity* associates the attribute value with its uncertainty and the uncertainty model used to characterize the uncertainty. The LDM only supports the storage and retrieving the uncertainty descriptors but does not handle uncertainties. It is the responsibility of the data processing functionalities to interpret and propagate the uncertainty information. It should be emphasized that typically the uncertainties are not constant quantities but depend on the quality of sensor readings, the availability of communication channels, eventual failure modes, etc. Consequently uncertainties should be propagated and handled in runtime in the data processing chain.

The LDM world model consists of static (e.g., road infrastructure elements) and dynamic (e.g., vehicles, foggy area) objects. Due to the high resolution of the map it is not feasible to enter the full static content into the LDM (see more about the onboard platform in the next section). Instead a “moving horizon” upload procedure is implemented (► Fig. 4.3). Depending on the location and intended driving direction of the vehicle, the relevant part of the static map is selected and loaded into the LDM dynamically. As the vehicle moves, static world objects are inserted/deleted to/from the LDM in order to maintain sufficient coverage of the world around. The static object stream is entered to the LDM via the situation assessment and coherency maintenance stages as “sensory input” because static map data is a highly trusted source and has crucial impact on the interpretation of other observations (e.g., position “calibration” via static landmarks).

The world model contains a rich, dynamic representation of the world: it supports building of dependable, robust applications by maintaining object and attribute uncertainties, object relations but leaves the responsibility of handling these to the applications. The next section overviews the guidelines for designing the “control side” of ► Fig. 4.1.

5 Situational Awareness in Control

The world model is maintained in order to use its content and generate “useful actions” for driver support, i.e., directly or indirectly influence the vehicle’s behavior. Consequently the action generating functionalities are called control functionalities – though they may not control the vehicles directly: the control loop may include the human driver in some applications. For example, an intelligent speed advice application can provide a speed set point on a display unit and leave the compliance to the driver or alternatively the speed advice application can directly set the cruising speed input of the adaptive cruise

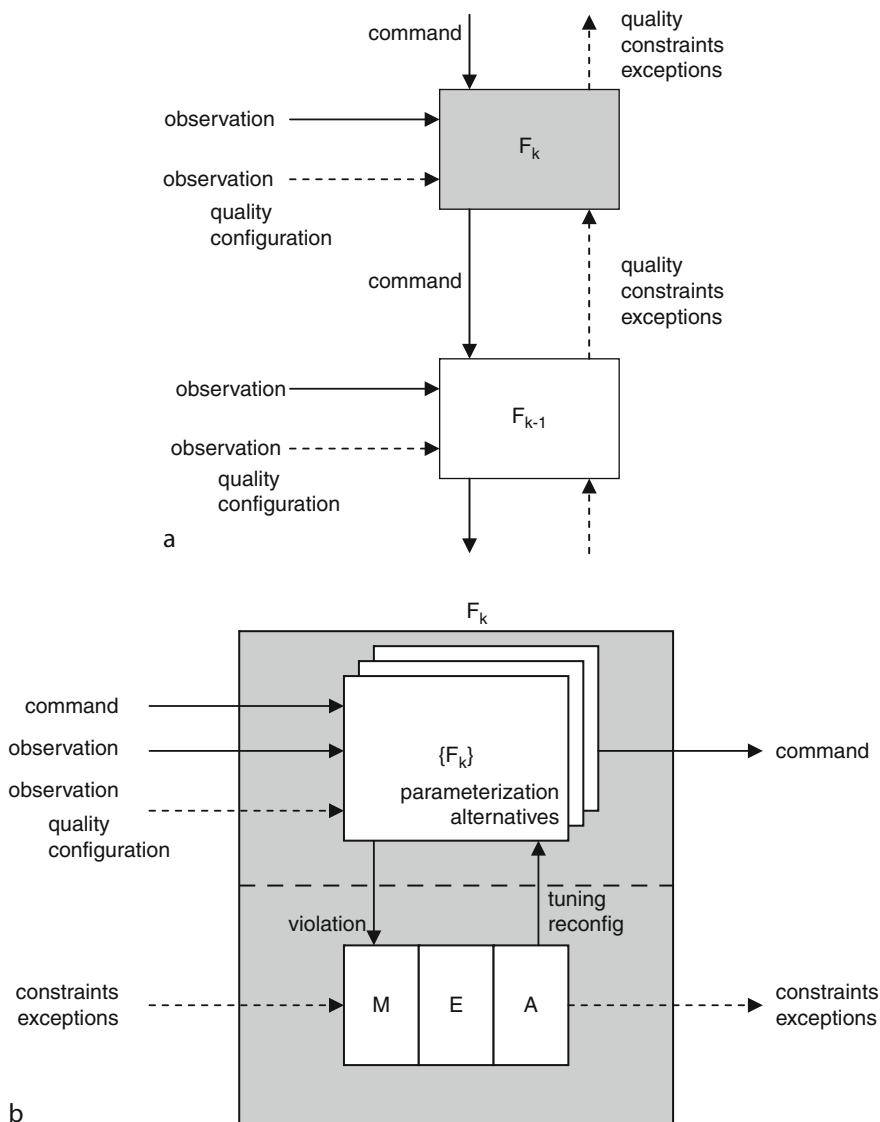
controller, i.e., leaving out the driver from the longitudinal control task. In complex control tasks, controllers act on different levels in the control hierarchy using the information maintained by the world model on different abstraction levels. Control architectures for real-time vehicle control is an intensively researched area and various proposals were published (Albus 2000; Horowitz and Varaiya 2000; Halle et al. 2004; Kester 2008; Urmson et al. 2008; Caveney 2010; Hurdus and Hong 2008). All these solutions are based on layered architectures: the control problem is formalized as hierarchy of control activities, each control layer working on a well-defined abstraction level. Each control layer receives observations on a matching abstraction level from the world model, i.e., reflecting on different findings in the situational awareness hierarchy.

► *Figure 4.6a* shows two consecutive control layers (F_k and F_{k-1}) in the control hierarchy.

The continuous arrows represent the “primary” signal flow. Commands arrive from the layer above (i.e., higher functional level, F_{k+1} for F_k), interpreting the observations available on the entity to be controlled, the controller calculates control commands for the lower layer (F_{k-1}). Besides this primary signal flow a secondary information flow (indicated by the dashed arrows) also has to be processed for the sake of situational awareness and thus dependable control. The control function takes into consideration the circumstances of the observations (in some cases it may include the health state of the vehicle platform itself) while generating the control command. For the commanding (upper) control layer it explicitly states what constraints should be maintained and the “quality” of the actions it can deliver under the current circumstances (which – in turn – should be taken into consideration when commands are generated). Achieving the state, what the incoming commands require, takes time and in this time window the circumstances can change dramatically – eventually violating the preconditions of the current control actions. These situations may result in exceptions indicating that the required control action cannot be completed. The exception is propagated upward in the control hierarchy: higher-level control layers have the ability to reconsider goals and initiate replanning.

► *Figure 4.6b* shows the conceptual processing scheme inside a control layer. In nontrivial cases a control functionality is realized via a set of parameterized alternatives. Each alternative can deliver the required functionality under well-defined circumstances called region of validity. Invoking an alternative outside its region of validity would result in incorrect operation. A parallel monitor-evaluate-act checks the conditions of the operation, handle exception and accordingly initiate parameter adjustment or reconfiguration in order to keep the operation within the region of validity. If this cannot be maintained it initiates aborting the operation and indicates an exception to the higher-level control layer.

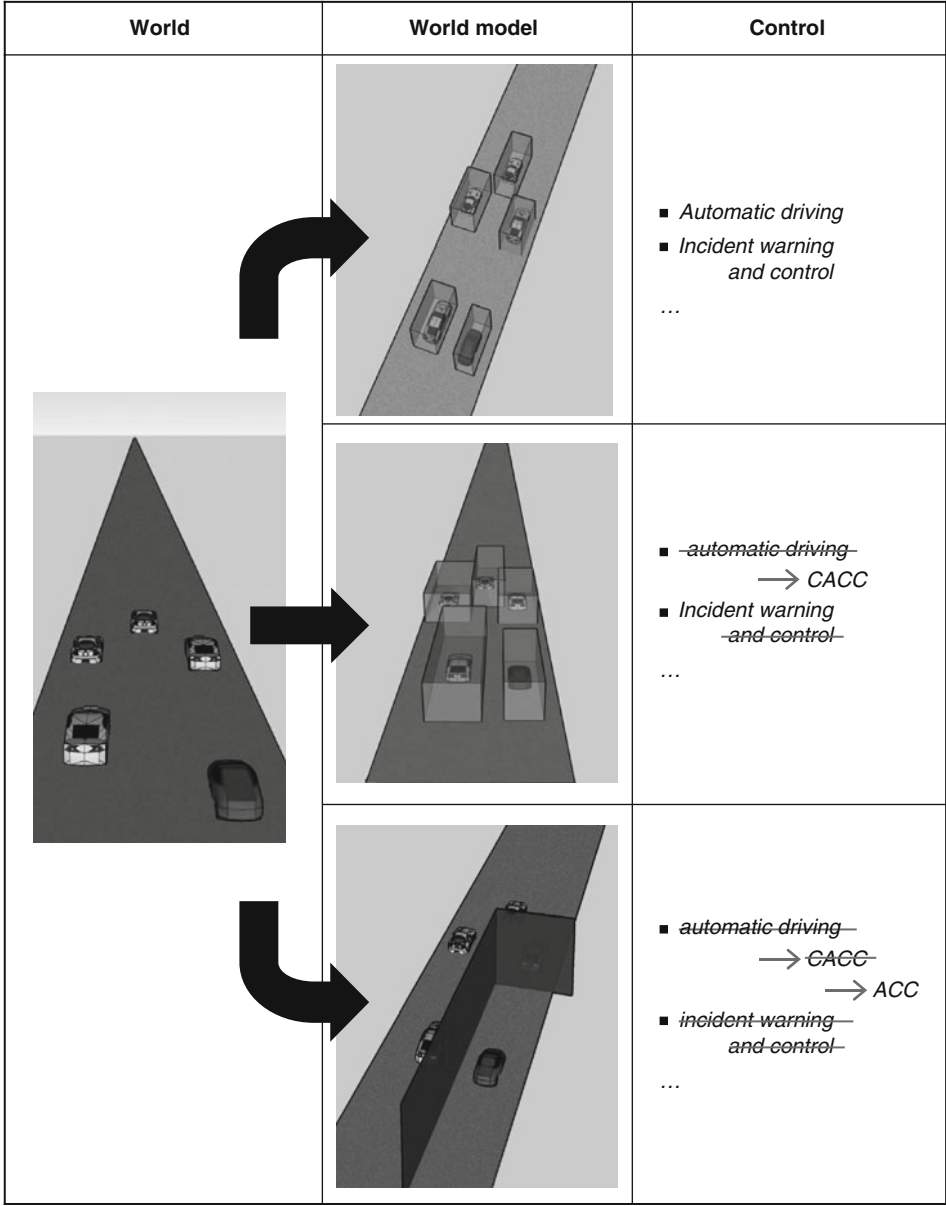
► *Figure 4.7* illustrates this type of interaction between situational awareness and control. Consider the highway driving configuration as shown in the left column in the table. Using onboard sensors, infrastructure based sensing, communication, etc., the red vehicle composes its internal world model (second column). Here only the uncertainties of vehicle localization and the uncertainty regions are considered and represented as semitransparent bounding boxes. Under optimal conditions the uncertainties are low,



■ Fig. 4.6

Layered control with situational awareness

enabling the execution of sophisticated control functionalities requiring accurate world representation, all relevant object relations maintained. If – due to, e.g., deteriorating visibility conditions, communication dropouts – the uncertainties become higher, the control system adjusts itself to the situation first by tuning control parameters to allow larger safety margins. If these measures are not enough certain automatic control functionalities should be dropped and, e.g., only the longitudinal control is kept by



■ Fig. 4.7
Example: situational awareness in longitudinal control

applying a cooperative adaptive cruise controller (CACC). If – because of failing communication – the vehicle can rely only on onboard sensors even the CACC functionality should be abandoned: a new alternative for longitudinal control, namely, the ACC controller is activated (van Arem et al. 2006).

In order to enable these decisions the conditions and constraints of the applicability of the data interpretation algorithms (including both on the sensory and the control side) and the responsibilities of these algorithms should be explicitly represented – i.e., characterizing all relevant data interpretation components with a “contract.” System level management functionalities should be incorporated, which monitor the validity of the contracts and have knowledge about handling violations. The spectrum of the system level management functionalities is wide. On one end of the spectrum there are systems, which simply shut down functionalities if the conditions of their applicability do not hold (i.e., it implements an “it’s better to do nothing than incorrect things” strategy); on the other end of the spectrum the management functionality adjusts the relevant subsystems to optimize the system level performance under particular operation conditions and availability of resources.

6 Conclusions

Intelligent vehicles are to perform sophisticated driver support and automated control functionalities in an unstructured, dynamic world full with uncertainties. On the other hand the margin of error is narrow: the resulting behavior should be safe under all conditions. Even if the driver is in charge the support system should feed her/him easily interpretable, valid information instead of generating confusion with misinterpreted information. The intelligent vehicle should understand the situation it is facing and should be able to adapt its internal functionalities accordingly.

The situation assessment process is not a “self-standing system component” but it is in the “fabric” of the intelligent vehicles software systems. It has far-reaching consequences. In the center point there is a modeling and representation issue: in order to describe, understand and manage a situation the conceptual model of the “world as we know it” should be built and maintained. The quality of the measured and derived quantities should be calculated, propagated along the signal processing chain as meta-data, and should be attached to the world model items they characterize. Situational awareness is a meaningless concept without control (automatic or human-in-the-loop type). The control/management/advice related functionalities should use the meta-data to adjust their operation accordingly assuring “doing the right thing at the right time.”

Beside the advance in computing and sensing technologies it is the ubiquitous availability of cheap communication, which accelerated the research and development on situational assessment. Observing and influencing large-scale distributed systems with complex dynamics (e.g., a traffic system) became a reality. Civil applications in mobility, environmental control, precision agriculture, etc., are emerging. A significant body of results can be “ported” from the defense industry, where the situational awareness problem surfaced earlier. Civil applications (including mobility) are more open. The deployment of large-scale sensor networks, participatory sensing, etc., will provide large, real-time and diverse data streams – which become only useful if their “meaning” can be decoded. The advance in the semantic web in general and the semantic sensor web in particular will bring the importance of the situation assessment to a new level.

References

- Albus JS (2000) 4-D/RCS reference model architecture for unmanned ground vehicles. Proceedings of the ICRA '00, (IEEE international conference on robotics and automation, 2000), San Francisco, CA, USA, 4:3260–3265
- Botts M et al (2007) OGC sensor web enablement: overview and high level architecture (OGC 07–165), Open Geospatial Consortium white paper. <http://www.opengeospatial.org/pressroom/papers>. Accessed 28 Dec 2007
- CAR 2 CAR Communication Consortium manifesto, car 2 car communication consortium. http://www.car-to-car.org/index.php?eID=tx_nawsecuredl&u=08&file=fileadmin/downloads/C2C-CC_manifesto_v1.1.pdf&t=1318038353&hash=aafc497a59410d1473eec47f8ca54e23d3f9d5a5. Accessed Oct 2011
- Caveney D (2010) Cooperative vehicular safety applications. *Control Syst IEEE* 30(4):38–53
- Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Hum Factors* 37(1):32–64
- Endsley MR (2000) Theoretical underpinning of situational awareness: a critical review. In: Endsley MR, Garland DJ (eds) *Situation awareness: analysis and measurement*. Lawrence Erlbaum Associates, Mahwah
- Hall DL, Llinas J (1997) An introduction to multisensor data fusion. *Proc IEEE* 85(1):6–23
- Halle S, Laumonier J, Chaib-Draa B (2004) A decentralized approach to collaborative driving coordination. Proceedings of the 7th international IEEE conference on intelligent transportation systems, Washington, DC, USA, pp. 453–458, 3–6 Oct 2004
- Horowitz R, Varaiya P (2000) Control design of an automated highway system. *Proc IEEE* 88(7):913–925
<http://www.safespot-eu.org/>
<http://www.w3.org/standards/semanticweb/>
- Hurdus JG, Hong DW (2008) Behavioral programming with hierarchy and parallelism in the DARPA urban challenge and robocup. Proceedings of the IEEE international conference on multisensor fusion and integration for intelligent systems, Seoul, South Korea, pp. 503–509, 20–22 Aug 2008
- Kester LJHM (2008) Designing networked adaptive interactive hybrid systems. Proceedings of the IEEE international conference on multisensor fusion and integration for intelligent systems, Seoul, South Korea, pp. 516–521, 20–22 Aug 2008
- Li L et al (2005a) IVS 05: new developments and research trends for intelligent vehicles. *Intell Syst* 20(4):10–14
- Li L, Wang F-Y, Kim H (2005) Cooperative driving and lane changing at blind crossings. In: Proceedings of the IEEE intelligent vehicles symposium, IEEE, Las Vegas, NV, USA, pp. 435–440
- Lygeros J, Godbole DN, Sastry S (1996) Multiagent hybrid system design using game theory and optimal control. In: Proceedings of the IEEE conference on decision and control, Kobe, 11–13 Dec 1996, pp. 1190–1195
- Lygeros J, Godbole DN, Sastry S (1998) Verified hybrid controllers for automated vehicles. *IEEE Trans Autom Control* 43(4):522–539
- Matheus CJ, Baclawski K, Kokar MM (2003) Derivation of ontological relations using formal methods in a situation awareness scenario. Proceedings of SPIE conference on multi-sensor, multi-source information fusion, Orlando, Cairns, Queensland, Australia, April 2003, pp. 298–309. Accessed Oct 2011
- Matheus CJ, Kokar M, Baclawski K (2003) A core ontology for situation awareness. Proceedings of the sixth international conference on information fusion 2003 1:545–552
- Michaud F, Lepage P, Frenette P, Letourneau D, Gaubert N (2006) Coordinated maneuvering of automated vehicles in platoons. *IEEE Trans Intell Transp Syst* 7(4):437–447
- NHTSA, crash avoidance metrics partnership. <http://www.nhtsa.gov/Research/Crash+Avoidance/Office+of+Crash+Avoidance+Research+Technical+Publications+2000-2010>. Accessed Oct 2011
- Papp Z, Brown C, Bartels C (2008) World modeling for cooperative intelligent vehicles. Proceedings of the IEEE intelligent vehicles symposium 2008, Eindhoven, The Netherlands, 4–6 June 2008, pp. 1050–1055
- Polychronopoulos A, Amditis A (2006) Revisiting JDL model for automotive safety applications: the PF2 functional model. 2006 9th international conference on Information fusion, Florence, Italy, 10–13 July 2006, pp. 1–7

- Schlenoff C, Washington R, Barbera T, Manteuffel C (2005) A standard intelligent system ontology. Proceedings of the unmanned ground vehicle technology VII conference (2005 SPIE defense and security symposium), Kissimmee, 28 Mar–1 Apr 2005
- Sheth A, Henson C, Sahoo SS (2008) Semantic sensor web. *Internet Comput IEEE* 12(4):78–83
- Steinberg AN, Bowman CL, White FE (1999) Revisions to the JDL data fusion model. Proceedings of the SPIE in sensor fusion. Architectures, algorithms, and applications, Orlando, FL, USA, vol. 3719
- Urmson C, Baker C, Dolan J, Rybski P, Salesky B, Whittaker W, Ferguson D, Darms M (2008) Autonomous driving in traffic: boss and the urban challenge. *AI Mag* 30(2):17–28
- van Arem B, van Driel CJG, Visser R (2006) The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans Intell Transp Syst* 7(4):429–436
- Varaiya P (1993) Smart cars on smart roads: problems of control. *IEEE Trans Autom Control* 38(2): 195–207
- Zott C, Yuen SY, Brown CL, Bartels C, Papp Z, Netten B (2008) Safespot local dynamic maps – Context-dependent view generation of a platform’s state & environment. Proceedings of the 15th world congress on intelligent transport systems, New York City, USA

5 Hierarchical, Intelligent and Automatic Controls

Bart De Schutter¹ · Jeroen Ploeg² · Lakshmi Dhevi Baskar¹ · Gerrit Naus³ · Henk Nijmeijer³

¹Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

²Automotive, TNO Technical Sciences, Helmond, The Netherlands

³Department of Mechanical Engineering, Dynamics and Control Section, Eindhoven University of Technology, Eindhoven, The Netherlands

1	<i>Introduction</i>	83
2	<i>Control Design Methods</i>	84
2.1	Static Feedback Control	84
2.2	Optimal Control and Model Predictive Control	85
2.2.1	Optimal Control	85
2.2.2	Model Predictive Control	86
2.3	Artificial Intelligence Techniques	88
3	<i>Hierarchical Control Framework</i>	89
3.1	PATH Framework	89
3.2	A Modified Version of the PATH Framework	91
3.3	Dolphin Framework	93
3.4	Auto21 CDS Framework	93
3.5	CVIS	94
3.6	SafeSpot	94
3.7	PREVENT	94
3.8	Comparison of IVHS Frameworks	95
4	<i>Control of Vehicle Platoons</i>	96
4.1	Problem Formulation	96
4.1.1	Vehicle Following	96
4.1.2	Control System Components	98
4.2	String Stability Literature Review	102
4.2.1	Platooning Requirements	102
4.2.2	The Lyapunov-like Approach	103
4.2.3	The Performance-Oriented Approach	103

4.2.4	The z-Domain Approach	105
4.2.5	Concluding Remarks on String Stability	106
4.3	Case Study	106
5	<i>Roadside and Area Control</i>	109
5.1	MPC for Roadside Controllers	109
5.2	MPC for Area Controllers	110
5.3	Interfaces Between the Different Control Layers	111
6	<i>Challenges and Open Issues</i>	112
7	<i>Conclusions</i>	113

Abstract: We present a survey on traffic management and control frameworks for Intelligent Vehicle Highway Systems (IVHS). First, we give a short overview of the main currently used traffic control methods that can be applied in IVHS. Next, various traffic management architectures for IVHS such as PATH, Dolphin, Auto21 CDS, etc., are briefly discussed and a comparison of the various frameworks is presented. Subsequently, we focus on control of vehicles inside a platoon, and we present a detailed discussion on the notion of string stability. Next, we consider higher-level control of platoons of vehicles. Finally, we present an outlook on open problems and topics for future research.

1 Introduction

Intelligent Vehicle Highway Systems (IVHS) (Sussman 1993; Fenton 1994; Bishop 2005) incorporate intelligence in both the roadway infrastructure and in the vehicles with the intention of reducing congestion and environmental impact, and of improving performance of the traffic network, by exploiting the distributed nature of the system and by making use of cooperation and coordination between the various vehicles and the various elements of the roadside infrastructure. IVHS comprise traffic management systems, driver information systems, and vehicle control systems. In particular, vehicle control systems are aimed at developing an automated vehicle highway system that shifts the driver tasks from the driver to the vehicle (Varaiya 1993). These driver tasks include activities such as steering, braking, and making control decisions about speeds and safe headways. Automated Highway Systems (AHS) go one step further than IVHS and involve complete automation of the driving task, with the vehicles being organized in platoons. For better (network-wide) coordination of traffic activities, AHS can also distribute the intelligence between the vehicles and the roadside infrastructure. In this chapter, we will focus on AHS and on the relations and interactions between the vehicles in the AHS inside platoons as well as with the roadside infrastructure. In particular, we will consider the control aspects of these systems.

The currently implemented traffic control and management systems are mainly using intelligence in the roadside infrastructure for controlling and managing the traffic system, However, such a system does not make use of the significant benefits offered by the intelligence – including the additional control, sensing, and communication capabilities – provided by (autonomous) Intelligent Vehicles (IVs). An interesting functionality that is allowed by full automation is to arrange the vehicles in closely spaced groups called “platoons” (Broucke and Varaiya 1997). To avoid collisions, intraplatoon spacing (i.e., vehicle spacing within a platoon) is kept very small and the interplatoon spacing is kept larger (Varaiya 1993; Li and Ioannou 2004). In the literature, many control frameworks, mainly intended to study intervehicle communication technologies and to control the platoon maneuvers in cooperation with the roadside infrastructure, have been developed and investigated (see, e.g., Hedrick et al. 1994; Rao and Varaiya 1994; Hsu et al. 1993; Tsugawa et al. 2001).

In this chapter we discuss various control methods and frameworks for platoons of IVs. After presenting a brief overview of the most frequent control methods that can be used for IVs, some general hierarchical frameworks for control of platoons of IVs are presented. Next, we consider control of vehicles inside platoons, with special attention to string stability. Afterward, control methods for the higher levels of the control hierarchy are presented. We conclude with an outlook and open issues.

2 Control Design Methods

In the literature different control methodologies have been presented for controlling and managing traffic networks (Papageorgiou 1983; Daganzo 1997; Kachroo and Özbay 1999). In this section we focus in particular on methods that also apply for platoons of IVs such as

- Static feedback control
- Optimal control and model predictive control
- Artificial intelligence (AI) techniques

The control methods we discuss below operate in discrete time. This means that at each sample time instant $t = kT$ where T is the sampling interval and the integer k is the discrete-time sample step, measurements of the traffic are performed and fed to the traffic controller. The controller then uses this information to determine the control signal to be applied during the next sampling interval.

2.1 Static Feedback Control

Dynamical systems can be controlled in two ways: using open-loop control and using closed-loop control. In an open-loop system, the control input does not depend on the output of the system, whereas in a closed-loop system, the control action is a function of the output of the system. Feedback or closed-loop control systems are particularly suited for applications that involve uncertainties or modeling errors.

In “static” (by “static” we mean here that the control parameters of the feedback controller are taken to be fixed) feedback control methods, the controller gets measurements from the system and determines control actions based on the current state of the system in such a way that the performance of the system is improved. The main examples of static feedback controllers are state feedback controllers and PID controllers (Åström and Wittenmark 1997).

The general form of a state feedback controller is $\mathbf{u}(k) = \mathbf{L}\mathbf{x}(k)$, where $\mathbf{x}(k)$ is the state vector at time step k , $\mathbf{u}(k)$ the (control) input to be applied at time step k , and \mathbf{L} is the feedback gain matrix. This feedback gain can be computed using, e.g., pole placement.

PID controllers are typically defined in continuous time and for single-input single-output systems, and they are of the form

$$u(t) = K_p \left(e(t) + \frac{1}{T_i} \int_0^t e(\tau) d\tau + T_d \frac{d}{dt} e(t) \right)$$

where $e(t)$ denotes the error signal between the measured output and the set-point value at time t , while the parameters K_p , T_i , and T_d denote, respectively, the proportional gain, the integral time constant, and the derivative time constant. To determine these parameters several tuning rules exist, such as the Ziegler–Nichols rules.

Static feedback strategies in general do not handle any external constraints. This is a major drawback of this control scheme.

2.2 Optimal Control and Model Predictive Control

Now we discuss two dynamic control methods that apply optimization algorithms to determine optimal control actions based on real-time measurements: optimal control and model predictive control.

2.2.1 Optimal Control

Optimal control determines a sequence of admissible control actions that optimize a performance function by considering future demands and by satisfying the constraints (Kirk 1970; Sussmann and Willems 1997). A general discrete-time optimal control problem contains the following elements:

- Dynamical system model equations
- An initial state \mathbf{x}_0
- An initial time t_0
- Constraints
- Measurements
- A performance index J

More specifically, consider a multi-input multi-output dynamical system expressed by the following equation:

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), \mathbf{d}(k))$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state vector, $\mathbf{u} \in \mathbb{R}^m$ the vector of manipulatable control inputs, \mathbf{f} is a continuously differentiable function, and \mathbf{d} is the disturbance vector.

For a given time horizon K , the optimal control problem consists in determining a sequence of control inputs $\mathbf{u}(0), \mathbf{u}(1), \dots, \mathbf{u}(K-1)$ in such a way that the performance

index J takes on the minimum possible value subject to the initial conditions, system dynamics, and constraints, i.e., minimize

$$J = \vartheta[\mathbf{x}(K)] + \sum_{k=0}^{K-1} \varphi(\mathbf{x}(k+1), \mathbf{u}(k), \mathbf{d}(k))$$

subject to


$$\begin{aligned} \mathbf{x}(0) &= \mathbf{x}_0 \\ \mathbf{x}(k+1) &= \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), \mathbf{d}(k)) \quad \text{for } k = 0, \dots, K-1, \\ \mathbf{u}_{\min}(k) &\leq \mathbf{u}(k) \leq \mathbf{u}_{\max}(k) \quad \text{for } k = 0, \dots, K-1, \\ \mathbf{c}(\mathbf{x}(k), \mathbf{u}(k), k) &\leq \mathbf{0} \quad \text{for } k = 0, \dots, K-1, \end{aligned}$$

where ϑ and φ are twice differentiable, nonlinear functions and are called the terminal cost and Lagrangian, respectively, \mathbf{u}_{\min} and \mathbf{u}_{\max} are bounds for the control variables, \mathbf{c} expresses path constraints imposed on the state \mathbf{x} and the control trajectories \mathbf{u} over the period $[t_0, t_0 + KT]$. The disturbance vector \mathbf{d} is assumed to be known over the period $[t_0, t_0 + KT]$. There are two basic approaches to solve the above optimal control problem: calculus of variations (Hestenes 1966; Gelfand and Fomin 1991) and dynamic programming (Bellman 1957).

The main drawback of optimal control is that the method is essentially an open-loop control approach and thus suffers from disturbances and model mismatch errors. Next, we will discuss model predictive control, which uses feedback and a receding horizon approach to overcome some of the drawbacks of optimal control.

2.2.2 Model Predictive Control

Model Predictive Control (MPC) (Maciejowski 2002; Rawlings and Mayne 2009) has originated in the process industry and it has already been successfully implemented in many industrial applications. MPC is a feedback control approach that can handle constrained, complex dynamical systems. The main difference between optimal control and MPC is the rolling horizon approach used in MPC (this essentially means that the optimal control is performed repeatedly but over a limited horizon). On the one hand, this results in a suboptimal performance compared to optimal control. However, on the other hand, the rolling horizon approach introduces a feedback mechanism, which allows to reduce the effects of possible disturbances and of model mismatch errors.

The underlying concept of the MPC controller (see  Fig. 5.1) is based on online optimization and uses an explicit prediction model to obtain the optimal values for the control measures subject to system dynamics and constraints. More specifically, at each control step k the state of the traffic system is measured or estimated, and an optimization is performed over the prediction period $[kT, (k + N_p)T]$ to determine the optimal control inputs, where N_p is the prediction horizon. Only the first value of the resulting control

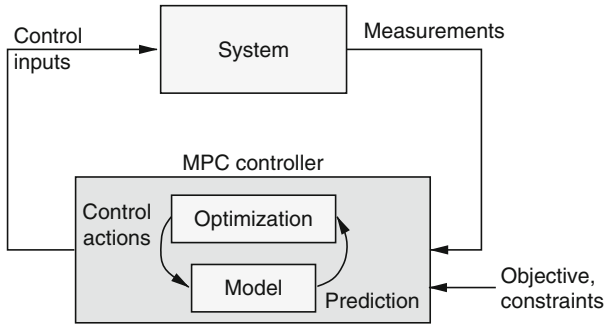


Fig. 5.1
Schematic view of the MPC approach

signal (the control signal for time step k) is then applied to the traffic system. At the next control step $k + 1$ this procedure is repeated.

To reduce the computational complexity and to improve stability often a control horizon N_c ($< N_p$) is introduced in MPC, and after the control horizon has been passed the control signal is taken to be constant.

There are two loops in MPC: the rolling horizon loop and the optimization loop inside the controller. The loop inside the controller of Fig. 5.1 is executed as many times as is needed to find (sufficiently) optimal control signals at control step k , for the given prediction horizon N_p , control horizon N_c , currently measured traffic state, and expected demands. The loop connecting the controller and the traffic system is performed once for each control step k and provides the state feedback to the controller. This feedback is necessary to correct for the ever present prediction errors, and to provide disturbance rejection (i.e., compensation for unexpected traffic demand variations). The advantage of this rolling horizon approach is that it results in an online adaptive control scheme that also allows us to take (slow) changes in the system or in the system parameters into account by regularly updating the model of the system.

MPC for linear systems subject to a quadratic objective function and linear constraints can be solved using quadratic programming. Other types of MPC problems in general require global or multi-start local optimization methods such as sequential quadratic programming, pattern search, or simulated annealing (Pardalos and Resende 2002).

Just as optimal control MPC can take into account constraints on the inputs and outputs, and it can also deal with multi-input multi-output systems. MPC has an advantage over optimal control due to receding horizon approach. This feedback mechanism of MPC makes the controlled system more robust to uncertainties and disturbances. Nevertheless, MPC still has some of the drawbacks of optimal control such as computational complexity, the need of an explicit model for prediction purposes, and the fact that the external inputs and disturbances need to be known fairly accurately in advance for the entire prediction horizon.

2.3 Artificial Intelligence Techniques

Artificial Intelligence (AI) techniques aim at imitating aspects of human intelligence and thinking while solving a problem by introducing human intelligence (to perceive a situation, to reason about it, and to act accordingly) into computer programs (Chen et al. 2008). AI techniques are mainly used in decision support systems, and the most important ones, in particular in the context of traffic control, are (Ritchie 1990; Sutton and Barto 1998; Nguyen and Walker 1999; Weiss 1999):

- Case-based reasoning
- Fuzzy logic
- Rule-based systems
- Artificial neural networks
- Multi-agent systems

Case-based reasoning, as the name suggests, solves a problem using the knowledge that was gained from previously experienced similar situations (cases) (Ritchie 1990; Aamodt and Plaza 1994). In this way, this technique learns the way a new problem is solved, tests the proposed solution using simulation methods, and stores the new solution in a database. A disadvantage of this approach is that it might not be clear what should be done for a case that is not yet present in the case base. However, new cases could be added online to deal with this problem.

Fuzzy logic systems, like humans, can handle situations where the available information about the system is vague or imprecise (Klir and Yuan 1995; Nguyen and Walker 1999). To deal with such situations, fuzzy sets are used to qualify the variables of the system in a non-quantitative way. Fuzzy sets are characterized using membership functions (e.g., Gaussian, triangle, or normal) that take a value between 0 and 1, and that indicate to what degree a given element belongs to the set (e.g., a speed could be 70% “high” and 30% “medium”). The membership degrees can then be used to combine various rules and to derive conclusions. This process consists of three parts: fuzzification, inference, and defuzzification. Fuzzification involves the transformation of a value of a variable into a fuzzy value, by linking it a given fuzzy set and determining a value for degree of membership. Inference uses a set of rules based on expert opinions and system knowledge and combines them using fuzzy set operators such as complement, intersection, and union of sets. Defuzzification converts the fuzzy output of the inference step in to a crisp value using techniques such as maximum, mean-of-maxima, and centroid defuzzification. One of the main difficulties of a fuzzy system can be the selection of appropriate membership functions for the input and output variables. Moreover, fuzzy systems are often combined with other AI techniques for their complete deployment.

Rule-based systems solve a problem using “if-then” rules (Hayes-Roth 1985; Russell and Norvig 2003). These rules are constructed using expert knowledge and stored in an inference engine. The inference engine has an internal memory that stores rules and information about the problem, a pattern matcher, and a rule applier. The pattern matcher searches through the memory to decide which rules are suitable for the

problem, and next the rule applier chooses the rule to apply. These systems are suited to solve problems where experts can make confident decisions. However, this system works only with already created rules and in its basic implementation it does not involve learning.

Artificial neural networks try to mimic the way in which the human brain processes information (Hammerstrom 1993; Yao 1999). These systems are useful in solving nonlinear problems where the rules or the algorithm to find solutions are difficult to derive. The basic processing unit of a neural network is called neuron or node. Each node fires a new signal when it receives a sufficiently high input signal from the other connected nodes. These nodes are organized in layers (an input layer, an output layer, and a number of hidden layers) and are interconnected by links or synapses, each associated with weights. A disadvantage is that artificial neural networks are non-informative models, and do not provide an explanation for the outcomes or for any failure that may occur in the process.

An agent is an entity that can perceive its environment through sensors and act upon its environment through actuators in such a way that the performance criteria are met (Ferber 1999; Weiss 1999). Multi-agent systems consist of a network of agents that are interacting among themselves to achieve specified goals. A high-level agent communication language is used by the agents for communication and negotiation purposes. Multi-agent systems can be applied to model complex systems, but their dynamic nature and the interactions between agents may give rise to conflicting goals or resource allocation problems.

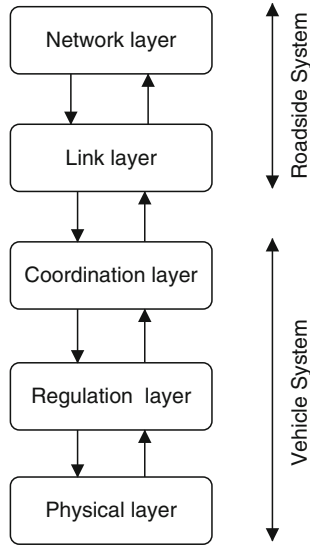
3 Hierarchical Control Framework

Now we discuss several control architectures that have been developed for linking the roadside infrastructure and automated platoons. In particular, we consider the PATH, Dolphin, Auto21 CDS, CVIS, SafeSpot, and PREVENT frameworks. We will in particular expand on the PATH framework as it will allow us to capture the control of platoons and collections of platoons later on in the chapter.

3.1 PATH Framework

The PATH architecture (Varaiya and Shladover 1991; Varaiya 1993; Broucke and Varaiya 1997; Horowitz and Varaiya 2000; Shladover 2007) mainly focuses on the coordination of roadside-vehicle and intervehicle activities.

The PATH framework considers a traffic network with many interconnected highways on which the vehicles are organized in platoons. The highways in the traffic network are considered to be divided into links (about 5 km long). A link is subdivided into segments (about 1 km long) with at least one exit or one entrance. A vehicle in the PATH framework is either considered as a leader, a follower, or a free agent (i.e., a one-vehicle platoon).



■ Fig. 5.2
PATH architecture

The PATH framework is a hierarchical structure in which the control of the automated highway system is distributed into five functional layers as shown in ► Fig. 5.2:

- Physical layer
- Regulation layer
- Coordination layer
- Link layer
- Network layer

The lower levels in this hierarchy deal with faster time scales (typically in the milli-seconds range for the physical layer up to the seconds range for the coordination layer), whereas for the higher-level layers the frequency of updating can range from few times per minute (for the link layer) to once every few minutes (for the network layer). The controllers in the physical, regulation, and coordination layer reside inside the vehicles. The physical and regulation controllers govern single vehicles, whereas the coordination layer involves several vehicles. The link layer and the network layer controllers are located at the roadside, with the link layer controllers managing single freeway segments, and the network layer controllers handling entire networks.

Now we discuss each layer of the PATH framework in more detail, starting from the bottom of the hierarchy:

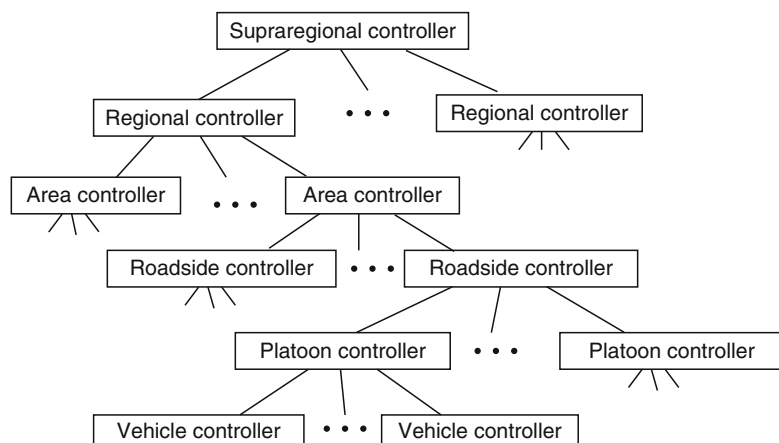
- The physical layer involves the actual dynamics of the vehicle. This layer has controllers that perform the actuation of the steering wheel, throttle, and brakes. It also contains the sensors in the vehicle that collect information about the speed, the acceleration, and the engine state of the vehicle, and send it to the regulation layer.

- The regulation layer controller executes the tasks specified by the higher-level coordination layer (such as lane changes, and splits or merges of platoons) by converting them into throttle, steering, and braking inputs for the actuators of the vehicle. The regulation layer controller within each vehicle uses feedback control laws to execute the lateral and longitudinal maneuvers and also notifies the coordination controller in case of any failures or unsafe outcomes of the maneuvers.
- The coordination layer receives its commands from the link layer (such as set-points or profiles for the speeds, or platoon sizes). A coordination layer controller allows coordination with other neighboring platoons using messages or communication protocols, and checks which maneuvers (like lane changes, splits, or merges) have to be performed by a vehicle in order to achieve the platoon size or path trajectory specified by the link controller.
- The link layer is mainly responsible for path and congestion control. Each link controller receives commands from the network layer (such as routes for the platoons) and based on these commands, the link controller calculates the maximum platoon size, and the optimum platoon velocity for each segment in the link it is managing. The link controller also sets the local path (which lane to follow) for each platoon.
- The network layer represents the top layer in the PATH hierarchy. At this layer, the controller computes control actions that optimize the entire network. Its task is to assign a route for each vehicle or platoon that enters the highway ensuring that the capacity of each potential route is utilized properly.

3.2 A Modified Version of the PATH Framework

In Baskar et al. (2007) and Baskar (2009), a modified version of the PATH framework was proposed. The main motivation for this modification is the following. Although the PATH framework includes both roadside infrastructure and vehicles, much of the research work was carried on the vehicle control side. In the PATH framework, the roadside controllers determine the activities that need to be carried out in different segments. However, when the platoon size is allowed to be long enough, it might be difficult for the roadside controller to assign the activities as the platoon resides in between two segments, and also for the vehicle controller to complete the activity within the specified space. For this reason, the modified framework uses platoon-based roadside controllers (so without segments). In addition, the framework also features some additional higher-level control layers.

The hierarchical control framework for IVHS proposed in Baskar et al. (2007) and Baskar (2009) is also based on the platoon concept and distributes the intelligence between the roadside infrastructure and the vehicles using control measures such as intelligent speed adaption, cooperative adaptive cruise control, lane allocation, on-ramp access control, route guidance, etc. The control architecture of Baskar et al. (2007) and Baskar (2009) consists of a multi-level structure with local controllers onboard the vehicles at the lowest level and one or more higher supervisory control levels, as shown in ► Fig. 5.3.



■ Fig. 5.3

The hierarchical control framework of Baskar et al. (2007) and Baskar (2009) for IVHS

The layers of the hierarchical control framework can be characterized as follows:

- The higher-level controllers (such as area, regional, and supraregional controllers) provide network-wide coordination of the lower-level and middle-level controllers. In particular, the area controllers provide area-wide dynamic route guidance for the platoons, and they supervise and coordinate the activities of the roadside controllers in their area by providing set-points, reference trajectories, or control targets. In turn, a group of area controllers can be supervised by regional controllers, and so on.
- The roadside controllers control a part of a highway or an entire highway. The main tasks of the roadside controllers are to assign speeds for each platoon, safe distances (to avoid collisions between platoons), and release times at the on-ramps. The roadside controllers also give instructions for merging, splitting, and lane changes to the platoons.
- The platoon controllers are responsible for control and coordination of each vehicle inside the platoon. These controllers are mainly concerned with actually executing the interplatoon maneuvers (such as merging with other platoons, splitting, and lane changing) and intraplatoon activities (such as maintaining safe intervehicle distances).
- The vehicle controllers present in each vehicle translate commands received from the platoon controllers (e.g., reference trajectories for speeds [for intelligent speed adaptation], headways [for cooperative adaptive cruise control], and paths) into control signals for the vehicle actuators such as throttle, braking, and steering actions.

Similar to the PATH framework, the lower levels in this hierarchy deal with faster time scales (typically in the milliseconds range for vehicle controllers up to the seconds range for roadside controllers), whereas for the higher-level layers the frequency of updating can range from few times per minute (for area controllers) to a few times per hour (for regional and supraregional controllers).

3.3 Dolphin Framework

The Japanese Dolphin framework developed by Tsugawa et al. (2000, 2001) is similar to the PATH architecture. The Dolphin framework considers vehicles to be arranged as platoons and develops intervehicle communication technologies to carry out cooperative driving for the purpose of smooth merging and lane changing.

The Dolphin framework is composed of three layers:

- Vehicle control layer
- Vehicle management layer
- Traffic control layer

The vehicle controller within each vehicle senses the states and the conditions ahead of the vehicle such as vehicle speed and acceleration and sends this information to the vehicle management layer. The vehicle controller also receives commands for the vehicle's steering actions and determines the actions for the vehicle actuators.

The vehicle management controller, which resides in each vehicle, receives suggestions for the movements of the vehicle from the traffic controller via road-vehicle communication and also considers the messages from the neighboring vehicles via intervehicle communication and the data received from the basic vehicle control layer. This controller determines the lateral and longitudinal movements of the individual vehicle under platoon-based driving.

The traffic control layer is common to all the vehicles and it is part of the roadside infrastructure. This layer consists of several distributed controllers, each of which determines advisory instructions for the vehicles in its own neighborhood and sends these instructions to the vehicle management layer.

3.4 Auto21 CDS Framework

The Auto21 Collaborative Driving System (CDS) framework (Hallé and Chaib-draa 2005) is mainly inspired by the concepts of the PATH and Dolphin architectures. The CDS architecture considers platoons of cars as autonomous agents and uses cooperative adaptive cruise control technologies to support platoon-based driving. The CDS framework employs an intervehicle coordination system that can ensure coordination of vehicle activities during their merge and split operations from a platoon and that can maintain stability among the vehicles in a platoon. The hierarchical architecture of the Auto21 CDS framework consists of the following three layers:

- Guidance layer
- Management layer
- Traffic control layer

with similar functionalities as the layers of the PATH and Dolphin architecture.

3.5 CVIS

CVIS (Cooperative Vehicle-Infrastructure Systems) (Toulminet et al. 2008; [CVIS Web site](#)) is a European research and development project that aims to design, develop, and test technologies that allow communication between the cars and with the roadside infrastructure, which in turn will improve road safety and efficiency, and reduce environmental impact.

CVIS operates with existing traffic control and management centers, roadside infrastructures, and vehicle systems. The complete system can be considered as a single-level architecture with the existing systems and CVIS operating at the same level. Various networks and communication protocols have been developed within CVIS to enable communication between different subsystems. The time scale for this architecture ranges from minutes to hours.

A CVIS system is composed of four subsystems: central, handheld, vehicle, and roadside subsystems. The central subsystem is basically a service provider for the vehicle or the roadside infrastructure. Typical examples of central subsystems include traffic control and service centers, and authority databases. The handheld subsystem provides services such as pedestrian safety and remote management of other CVIS subsystems by allowing access to the CVIS system using PDAs and mobile phones. The vehicle subsystem is comprised of onboard systems and includes vehicle sensors and actuators, and equipment for vehicle-vehicle and vehicle-infrastructure communication. The roadside subsystem corresponds to the intelligent infrastructure that operates at the roadside and includes traffic signals, cameras, variable message signs, etc.

3.6 SafeSpot

SafeSpot (Toulminet et al. 2008; [SafeSpot Web site](#)) is a research project funded by the European 6th Framework Program on Information Society Technologies. The main objective of this project is to improve road safety using advanced driving assistance systems and intelligent roads. The safety margin assistant developed by the SafeSpot project uses advanced communication technologies to obtain information about the surrounding vehicles and about the roadside infrastructure. This safety margin assistant can detect dangerous situations in advance and can make the driver aware of the surrounding environment using a human machine interface. The time scale for this architecture ranges from seconds to minutes.

3.7 PReVENT

PReVENT ([PReVENT Web site](#)) is a European automotive industry activity co-funded by the European Commission. The main focus of the PReVENT project is to develop preventive applications and technologies that can improve the road safety. These safety

applications use in-vehicle systems to maintain safe speeds and distances depending on the nature and severity of the obstacles, and to provide instructions and to assist the drivers in their driving tasks so as to avoid collisions and accidents.

The PReVENT architecture features a three-layer approach with the following layers: perception layer, decision layer, and action layer. All these layers are located within the vehicle. From the perception layer upward to the action layer, the time complexity and update frequency of states typically ranges from milliseconds to seconds.

The perception layer uses onboard sensors (such as radar, cameras, and GPS receivers) in conjunction with digital maps and allows vehicle-to-vehicle and vehicle-to-infrastructure communication. The decision layer assesses dangerous situations ahead of the vehicle and determines relevant actions that are needed to avoid such situations. The controller then passes this decision to the action layer. The action layer then issues warnings to the driver about the severity of the situation through an appropriate human machine interface or through vehicle actuators such as the steering wheel or the brakes.

3.8 Comparison of IVHS Frameworks

The main differences between the frameworks consist in the control objectives considered, the type of formation control (platoons or single cars), the location of the intelligence (roadside and/or in-vehicle), and communication and coordination mechanism.

The PATH, Dolphin, AUTO21 CDS, and CVIS frameworks have developed control methodologies to be implemented in the roadside infrastructure to improve the traffic flow or in vehicles to allow automation of driving tasks. On the other hand, SafeSpot and PReVENT focus on improving the road safety by avoiding or preventing accidents, and they aim at integrated safety, with an emphasis on the potential of communication between vehicles and between vehicles and roadside systems.

The frameworks usually consider the vehicles to be controlled either as part of higher-level entities such as platoons, or as individual vehicles. The PATH, Dolphin, and Auto21 CDS frameworks allow platooning. On the other hand, SafeSpot, PReVENT, and CVIS do not use platoons.

The PATH framework allows involvement of both roadside infrastructure and vehicles for improving traffic performance. Although the Dolphin and the Auto21 CDS frameworks consider distributed intelligence between roadside infrastructure and vehicles, the roadside infrastructure only provides suggestions and instructions to the vehicles. The platoons are not obliged to follow these suggestions. CVIS and SafeSpot incorporate intelligence in both vehicle and roadside infrastructure. PReVENT also includes distributed intelligence but with the main focus on vehicle intelligence.

Almost all the frameworks and projects have designed and developed technologies for intervehicle and roadside-vehicle communication for coordination of activities. More specifically, PATH has developed dedicated communication protocols and the Dolphin framework has developed a wireless local access network model for vehicle following, and intervehicle communication technologies for coordination of platoon maneuvers.

For the coordination of tasks within the platoons, the AUTO21 CDS framework allows both a centralized setup (i.e., the platoon leader instructs intraplatoon maneuvers) and a decentralized setup (i.e., all the members of the platoon are involved in the coordination). SafeSpot, PReVENT, and CVIS focus on the issue of developing communication techniques that can be implemented in existing traffic networks and that can also be extended to AHS.

A more detailed comparison of the above frameworks is presented in Baskar (2009) and Baskar et al. (2011).

4 Control of Vehicle Platoons

The traffic control frameworks presented in the previous section either explicitly identify a vehicle platoon as a layer in the hierarchical structure (PATH, Dolphin, Auto21 CDS) or have a more generic design, such that vehicle platoons can be an important component in the overall control framework (CVIS, PReVENT). Consequently, vehicle platoons are expected to play an important role in traffic control. Therefore, this section focuses on the developments in the field of control of vehicle platoons. To this end, we will first formally introduce the platoon control problem after which a very important aspect, being the notion of “string stability” is reviewed.

4.1 Problem Formulation

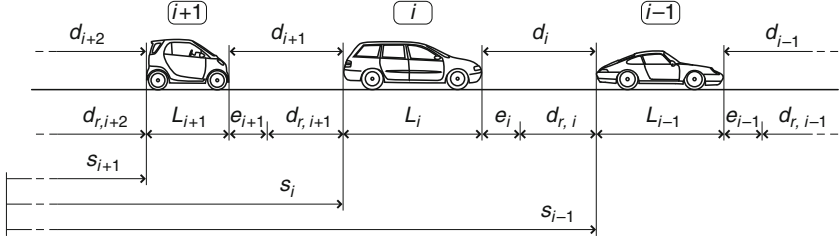
The concept of platooning refers to a string of vehicles that aim to keep a specified, but not necessarily constant, intervehicle distance. Consequently, the control of vehicles in a platoon can be characterized as a vehicle-following control problem. This section will state the formal control objective and identifies the components of the control system, as reported in literature.

4.1.1 Vehicle Following

Let us consider an arbitrary (and possibly infinite) number of vehicles that drive behind each other as depicted in Fig. 5.4. Note that this configuration is referred to as a “string” of vehicles and not as a “platoon” because the latter might suggest the presence of a platoon leader, which is certainly not a prerequisite.

In Fig. 5.4, the position (measured with respect to the rear bumper) of vehicle i at time t is denoted by $s_i(t)$. The distance (“headway”) $d_i(t)$ of vehicle i , with length L_i , at time t to the preceding vehicle $i - 1$ therefore equals

$$d_i(t) = s_{i-1}(t) - (s_i(t) + L_i). \quad (5.1)$$



■ Fig. 5.4
A string of vehicles on a straight lane

The distance error $e_i(t)$ is taken as

$$e_i(t) = d_i(t) - d_{r,i}(t) \quad (5.2)$$

where $d_{r,i}(t)$ is the desired headway of vehicle i that follows from the so-called spacing policy (see below). For the ease of notation we will not explicitly indicate the time argument t any more in the sequel if it is clear from the context.

The control objective \mathcal{O} , referred to as (asymptotic) vehicle following, can now be formulated in its most basic form as regulating e_i to zero in the presence of disturbances induced by:

- Initial condition errors of any vehicle in the string
- Perturbations in the velocity of other vehicles in the string
- Velocity variations of the lead vehicle
- Set-point changes with respect to the desired distance, or, in other words

$$\mathcal{O} : \lim_{t \rightarrow \infty} e_i(t) = 0, \quad \forall t \geq t_0, \quad \forall i \in \{2, \dots, m\} \quad (5.3)$$

where control starts at time $t = t_0$. Here, the string is assumed to consist of m vehicles in total, of which $m - 1$ vehicles have a vehicle-following objective. The lead vehicle, with index 1, not being subject to a vehicle-following objective, ultimately defines the desired speed of the entire string. To this end, the lead vehicle is assumed to be velocity controlled with a given, possibly time-varying set speed. For persistently time-varying perturbations in the velocity of other vehicles in the string and/or of the lead vehicle, which must certainly be included for the vehicle-following problem, it might not be feasible to achieve asymptotic disturbance rejection or tracking, respectively. In such cases, the control objective can be formulated as

$$\mathcal{O} : \|e_i(t)\|_p \leq \varepsilon, \quad \forall t \geq T, \quad \forall i \in \{2, \dots, m\} \quad (5.4)$$

where ε is an a priori chosen (small) number, and $T > t_0$ is introduced to ignore transient effects due to initial condition errors. In ► Eq. 5.4 $\|\cdot\|_p$ denotes the vector p -norm, defined as

$$\|\mathbf{u}\|_p \triangleq \left(\sum_i |u_i|^p \right)^{1/p} \quad (5.5)$$

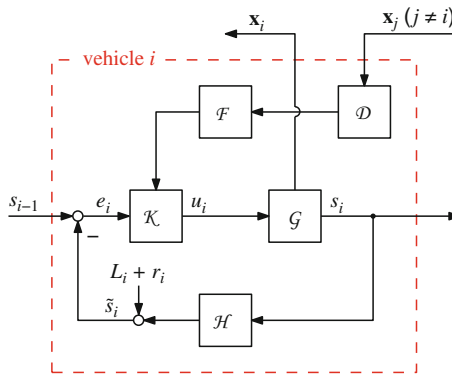
for a vector \mathbf{u} . Whichever objective is used, the vehicle-following problem can essentially be regarded as a standard asymptotic tracking problem that occurs in many applications. One additional requirement that can be formulated, however, is related to how the responses of the vehicles to disturbances evolve not only in time, but also across the vehicles in the string. This disturbance evolution across a string of vehicles, or, in general, across a number of interconnected subsystems, is covered by the notion of *string stability*. Note that the problem formulation can be further extended with additional performance requirements based on comfort, fuel consumption, and other important criteria. These are however outside the scope of this section.

4.1.2 Control System Components

In order to further explain the various components in the control system that is designed to fulfill the objective \blacktriangleright Eq. 5.3 or \blacktriangleright Eq. 5.4, \blacktriangleright Fig. 5.5 shows a block scheme of a controlled vehicle in a string, taken from Naus et al. (2009, 2010c). Although this block scheme will not hold for all possible control solutions, it serves the purpose of identifying the various components, to be explained in further detail below.

First of all, the subsystem \mathcal{G} represents the vehicle to be controlled. In frequency-domain-oriented approaches, such as applied in Naus et al. (2009, 2010c), a general linear vehicle model is formulated, according to

$$\mathcal{G} : G_i = \frac{k_i}{s^2(\tau_i s + 1)} e^{-\theta_i s} \quad (5.6)$$



■ Fig. 5.5

General control structure of the vehicle-following problem

where τ_i is a time constant representing the lumped vehicle actuator dynamics, θ_i is a time delay, caused by the throttle and brake system, and k_i a gain. The input u_i can be interpreted as the desired acceleration, whereas the output is the resulting vehicle position s_i (i.e., $k_i \approx 1$). Consequently, the model of [Eq. 5.6](#) includes in fact a low-level acceleration controller. This acceleration controller can be characterized as a linearizing pre-compensator whose input is the desired acceleration and whose output is the throttle valve position or brake pressure. This pre-compensator aims to linearize the vehicle drive line at the lowest level possible, thereby greatly simplifying the design of higher-level vehicle-following controllers. Moreover, the pre-compensator “hides” specific vehicle characteristics such as the mass, aerodynamic drag, and rolling resistance, simplifying even more the higher-level control design. The following parameter values for a Citroën Grand C4 Picasso combined with a custom design pre-compensator are mentioned in Naus et al. (2010c): $k_i = 0.72$, $\tau_i = 0.38$ s and $\theta_i = 0.18$ s.

Another frequently used vehicle model is obtained by means of input–output linearization by state feedback (Stanković et al. 2000), resulting in

$$\mathcal{G} : \begin{cases} \dot{s}_i = v_i \\ \dot{v}_i = a_i \\ \dot{a}_i = -\tau_i^{-1}(v_i)a_i + \tau_i^{-1}(v_i)u_i \end{cases} \quad (5.7)$$

where v_i is the vehicle speed, a_i the acceleration, and u_i the external input (desired acceleration); $\tau_i(v_i)$ is a velocity-dependent time constant representing the engine dynamics. For ease of notation, the time argument t is omitted. Note that [Eq. 5.7](#) is in fact only partly linearized since the state v_i still occurs in a nonlinear fashion. Obviously, it is easily possible to complete the model linearization by introducing a new input η_i and choosing $u_i = \tau_i(v_i)\eta_i + a_i$, resulting in

$$\mathcal{G} : \begin{cases} \dot{s}_i = v_i \\ \dot{v}_i = a_i \\ \dot{a}_i = \eta_i \end{cases} \quad (5.8)$$

which is applied in, e.g., Sheikholeslam and Desoer (1992), Ioannou and Chien (1993), and Sheikholeslam and Desoer (1993). Alternatively, τ_i could be approximated by a constant value. The model of [Eq. 5.7](#) then becomes linear and equal to the frequency-domain model of [Eq. 5.6](#) with $k_i = 1$ and $\theta_i = 0$.

The subsystem \mathcal{H} describes the spacing policy, which refers to the choice of desired headway $d_{r,i}$. Originally, $d_{r,i}$ has been determined using physical considerations (Ioannou and Chien 1993), taking into account the distance it takes for a vehicle to adapt its speed to a preceding decelerating vehicle, such that a collision is just avoided. This approach results in a desired headway according to

$$d_{r,i} = c_{0,i} + c_{1,i}v_i + c_{2,i}(v_{i-1}^2 - v_i^2) \quad (5.9)$$

where the coefficients $c_{k,i}$ ($k = 0, 1, 2$) depend on vehicle specifications, such as maximum deceleration and jerk, and a possible reaction time delay. Obviously, for tight vehicle

following, i.e., v_i close to v_{i-1} , ⑤ Eq. 5.9 can be simplified to $d_{r,i} = c_{0,i} + c_{1,i}v_i$, which is commonly denoted as

$$d_{r,i} = r_i + h_i v_i. \quad (5.10)$$

The variable h_i is known as the (desired) time headway and r_i is the standstill distance, since it can be interpreted as such. The latter is desired to prevent a near collision at standstill. Note that according to International Organization for Standardization (2002), h_i must be greater than or equal to 1.0 s for commercially available road vehicles equipped with Adaptive Cruise Control (ACC). (The letter “A” in ACC turns out to be rather versatile, as it might represent “Adaptive,” “Automatic,” “Autonomous,” “Advanced,” or “Active,” depending on the car brand). There is no legal upper limit to h_i , but in practice this turns out to be chosen as high as 3.6 s by system manufacturers, probably stemming from comfort requirements. This is significantly higher than common human driver behavior would incur. For Cooperative ACC (see below), string stability can be achieved for values of h_i smaller than 1.0 s (Naus et al. 2009, 2010c); safety is however not taken into account here. The spacing policy ⑤ Eq. 5.10 can be formulated in terms of a transfer function as follows. First write the distance error e_i as defined in ⑤ Eq. 5.2 as

$$e_i = d_i - d_{r,i} = s_{i-1} - (s_i + L_i) - (r_i + h_i v_i) = s_{i-1} - \tilde{s}_i \quad (5.11)$$

with

$$\tilde{s}_i = L_i + r_i + s_i + h_i v_i. \quad (5.12)$$

Note that \tilde{s}_i can be interpreted as the “virtual control point” of vehicle i , whose position must be as close as possible to the actual position s_{i-1} of the preceding vehicle $i-1$. Reformulating ⑤ Eq. 5.12 as a system with input s_i and output \tilde{s}_i , while omitting the constants L_i and r_i , yields the spacing policy transfer function

$$\mathcal{H} : H_i(s) = 1 + h_i s \quad (5.13)$$

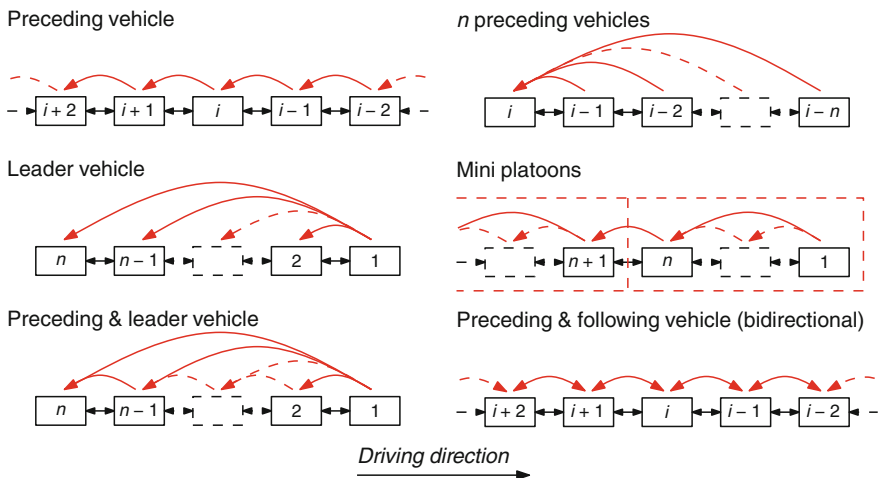
which clearly indicates the differential action contributing to a well-damped behavior of the controlled vehicle. This explains why this spacing policy is widely used in literature and has become a de facto standard for commercial ACC. Note that the chosen spacing policy also influences the traffic flow stability, having consequences for throughput. These aspects are not taken into account, even though it is known that the above constant time headway spacing policy might have adverse effects on traffic flow stability. Some research has been done into this area; see, e.g., Swaroop and Rajagopal (1999), who analyze the effects of the spacing policy on traffic flow stability and propose alternative spacing policies.

The feedback controller \mathcal{K} in the block diagram of ⑤ Fig. 5.5 can be regarded as an ACC controller, based on direct measurement of the distance to the preceding vehicle. In general, this distance is measured by an onboard environmental sensor such as radar or scanning laser (lidar). (A radar or laser directly measures the distance d_i . The block diagram is, however, rather efficiently formulated due to the introduction of the spacing policy \mathcal{H} , having the drawback that d_i is not explicitly available in the diagram anymore.)

PD-like controllers are often used here, see, e.g., Ioannou and Chien (1993) and Venhovens et al. (2000), where the differential action does not have to be explicitly implemented since both radar and lidar directly measure the relative velocity. In Naus et al. (2010a), an ACC controller is designed based on Explicit MPC, which is a rather promising approach in view of constraints such as maximum velocity, acceleration, and jerk.

The feedforward controller, denoted by \mathcal{F} , provides the extension from ACC to so-called Cooperative ACC (CACC). To this end, wireless communication is applied to obtain motion data, represented by the vehicle state \mathbf{x}_j in \bullet Fig. 5.5, from other vehicles than the directly preceding one, and/or to obtain data \mathbf{x}_{i-1} from the directly preceding vehicle that cannot be measured by the onboard environment sensor. In literature, a very wide range of communication structures is found, e.g., communication with the directly preceding vehicle, communication with a platoon leader, or even bidirectional communication. \bullet Figure 5.6 provides an overview of structures reported in literature.

The main objective here is to obtain or enhance string stability compared to ACC. The (possibly varying) communication delay involved in this method of data exchange is represented by \mathcal{D} . The combined controllers \mathcal{K} and \mathcal{F} constitute a CACC controller. In the literature, a PD-like controller with feedforward is very often used (see, e.g., Sheikholeslam and Desoer 1992, 1993; Tsugawa et al. 2001; Naus et al. 2009, 2010c). Also sliding mode controllers are regularly encountered in literature (see, e.g., Swaroop et al. 1994; Gehring and Fritz 1997). Note that Levine and Athans (1966) is probably the first paper on the subject. In this paper, optimal control is applied, which is not surprising given the developments at the time. Although \mathcal{K} and \mathcal{F} obviously need to be synthesized in an integral approach, and usually are, it is still possible to distinguish both functions in the controller and even to guarantee string stability in case only the “ACC-part” \mathcal{K} is



■ Fig. 5.6
Communication structures for CACC

active, i.e., when the communication link is not functioning properly, albeit with a significantly larger time headway. The latter feature might prove of great importance to obtain a certain robustness against communication impairments such as latency (caused by among others queuing delay, transmission delay, and propagation delay) and packet loss (due to packet collisions related to the so-called hidden-node problem, and interference).

4.2 String Stability Literature Review

This section provides a short literature overview on the notion of string stability. Three main approaches can be distinguished here, being a Lyapunov-like approach, focusing on generalization of the notion of stability, a performance-oriented approach, and finally a linear stability approach for strings of infinite length, based on the bilateral \mathcal{Z} -transform.

4.2.1 Platooning Requirements

From a practical perspective, control design for a string of vehicles in order to obtain vehicle-following behavior entails in the first place achieving *individual vehicle stability* (Rajamani 2006). This is commonly interpreted as achieving stable vehicle-following behavior with the preceding vehicle driving at constant velocity. Taking the block diagram of Fig. 5.5, with vehicle model $G_i(s)$, controller $K_i(s)$, and spacing policy $H_i(s)$, the system with complementary transfer function $T_i(s) = (1 + G_i(s)K_i(s)H_i(s))^{-1} G_i(s)K_i(s)$ should be stable. Obviously, this is a basic practical requirement that is assumed to be met in the remainder of this section.

Besides individual vehicle stability, an important requirement is the ability of the string of vehicles to attenuate disturbances, or at least guarantee boundedness, introduced by an arbitrary vehicle in the string as we “move away” from that vehicle. Assume for instance a vehicle string where each vehicle is controlled based on information of one or more preceding vehicles, i.e., a unidirectional communication link between vehicles in upstream direction. If the first vehicle in the string introduces some disturbance, e.g., a variation in velocity, then the states of the following vehicles should be bounded as a result or, preferably, get smaller in some sense in upstream direction, ultimately leading to a smooth traffic flow. The notion of string stability refers exactly to this property. Note that Swaroop et al. (1994) also mention that spacing errors should not be amplified in the platoon in order to avoid collisions, thereby providing another motivation for string stability. (It might be questioned whether collision avoidance is a valid argument for requiring string stability, since this is likely to require dedicated controllers that do not aim to optimize string behavior with respect to smoothness. Moreover, collision avoidance cannot be guaranteed with linear controllers, as regularly used in controller design for vehicle platoons.)

4.2.2 The Lyapunov-like Approach

Although in the majority of the literature string stability is not explicitly defined, a formal approach of the subject can be found in the work of Swaroop (Swaroop and Hedrick 1996; Swaroop et al. 2001). As opposed to system stability, which is essentially concerned with the evolution of system states over time, string stability focuses on the propagation of states over subsystems. Recently, new results appeared (Klinge and Middleton 2009), related to the stability analysis in case of a one-vehicle look-ahead control architecture and a homogeneous string. The resulting string stability definition is given below.

Definition 4.1

(\mathcal{L}_p -String Stability). Consider a string of m dynamic systems of order n described by

$$\begin{aligned}\dot{\mathbf{x}}_1 &= \mathbf{f}(\mathbf{x}_1, \mathbf{0}) \\ \dot{\mathbf{x}}_i &= \mathbf{f}(\mathbf{x}_i, \mathbf{x}_{i-1}) \quad \forall i \in \{2, \dots, m\}\end{aligned}$$

with $\mathbf{x}_i \in \mathbb{R}^n$, $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, and $\mathbf{f}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$, and $\mathbf{x}_i(0) = \mathbf{0}$ for $i = 2, \dots, m$. Then, the origin is \mathcal{L}_p -string stable if for each $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\|\mathbf{x}_1(0)\|_p < \delta \Rightarrow \sup_i \|\mathbf{x}_i(t)\|_{\mathcal{L}_p} < \varepsilon \quad (5.14)$$

Here, $\|\cdot\|_p$ denotes the vector p -norm and $\|\cdot\|_{\mathcal{L}_p}$ denotes the \mathcal{L}_p -norm for vector-valued signals. Obviously, according to this definition, homogeneous, linear strings of finite length are string stable, provided that the vehicles are individually stable. However, as m approaches infinity, it appears that the string stability requirement leads to a lower bound on the time headway h . The above definition nicely illustrates that string stability is concerned with the propagation of states over the string.

4.2.3 The Performance-Oriented Approach

Despite the existence of the Lyapunov-like approach, which may be thought of as being rigorous, a frequency-domain approach for string stability is also adopted since this appears to directly offer tools for controller synthesis (Sheikholeslam and Desoer 1992, 1993; Ioannou and Chien 1993; Swaroop et al. 1994; Gehring and Fritz 1997; Stanković et al. 2000; Naus et al. 2009, 2010c). In the performance-oriented approach, string stability is evaluated by analyzing the amplification in upstream direction of either distance error, velocity, and/or acceleration. This immediately leads to the following definition, used (implicitly) in the above literature references.

Definition 4.2

(Bounded Propagation String Stability). Consider a string of $m \in \mathbb{N}$ dynamic systems, then the string is considered string stable if and only if

$$\|z_i(t)\|_{\mathcal{L}_\infty} \leq \|z_{i-1}(t)\|_{\mathcal{L}_\infty}, \quad \forall i \in \{2, \dots, m\}, t \geq 0$$

where $z_i(t)$ can either be the distance error $e_i(t)$, the velocity $v_i(t)$, or the acceleration $a_i(t)$ of vehicle i , $z_1(t) \in \mathcal{L}_\infty$ is any given input signal, and $z_i(0) = 0$ for $i = 2, \dots, m$.

Here, $\|\cdot\|_{\mathcal{L}_\infty}$ denotes the signal ∞ -norm, which for scalar signals comes down to taking the highest peak value over time. Definition 4.2 thus states that the peak value of either distance error, velocity, or acceleration must decrease in upstream direction. In literature, the choice between distance error, velocity, or acceleration seems a little arbitrary. Note that $z_i(t)$ is thus assumed to be a scalar signal.

From linear system theory (Desoer and Vidyasagar 2009), it is well known that the \mathcal{L}_∞ -norms of input and output are related through the \mathcal{L}_1 -norm of the impulse response matrix $\gamma_i(t)$ with respect to the “input” z_{i-1} and the “output” $z_i(t)$, according to

$$\|\gamma_i(t)\|_{\mathcal{L}_1} = \max_{z_{i-1} \neq 0} \frac{\|z_i(t)\|_{\mathcal{L}_\infty}}{\|z_{i-1}(t)\|_{\mathcal{L}_\infty}} \quad (5.15)$$

where the \mathcal{L}_1 -norm $\|\gamma_i(t)\|_{\mathcal{L}_1}$ for scalar impulse responses equals the integral over time of the absolute value $|\gamma_i(t)|$. Hence, the above definition yields the *necessary* and *sufficient* string stability requirement (Rajamani 2006)

$$\|\gamma_i(t)\|_{\mathcal{L}_1} \leq 1, \quad \forall i \in \{2, \dots, m\}, \quad t \geq 0. \quad (5.16)$$

In practice, however, the application of impulse response functions is not particularly easy and intuitive. Using another fact from linear system theory allows translation of the string stability requirement to the frequency domain, to some extent. Consider to this end the relation between the norm of $\gamma_i(t)$ and its corresponding transfer function $\Gamma_i(s)$ on the imaginary axis:

$$\|\Gamma_i(j\omega)\|_{\mathcal{H}_\infty} \leq \|\gamma_i(t)\|_{\mathcal{L}_\infty} \quad (5.17)$$

where the \mathcal{H}_∞ norm for scalar transfer functions equals the peak over the frequency ω of the gain $|\Gamma_i(j\omega)|$. This immediately leads to the *necessary* condition for string stability


$$\|\Gamma_i(j\omega)\|_{\mathcal{H}_\infty} \leq 1, \quad \forall i \in \{2, \dots, m\} \quad (5.18)$$

which is far more convenient for controller synthesis in general. In Swaroop et al. (1994), it is shown that if $\gamma_i(t) \geq 0$, \blacklozenge Eq. 5.17 becomes an equality, thus making \blacklozenge Eq. 5.18 a necessary and sufficient condition. From a physical perspective, $\gamma_i(t) \geq 0$ means that the time response shows no overshoot to a step input, indicating a sufficient level of damping. Commonly, the requirement of \blacklozenge Eq. 5.18 is considered necessary and sufficient for string stability. This can be motivated by the fact that the \mathcal{H}_∞ -norm is induced by the \mathcal{L}_2 -norms of input and output which, in turn, are measures for energy. As a consequence, the condition $\|\Gamma_i(j\omega)\|_{\mathcal{H}_\infty} \leq 1$ can be interpreted as requiring energy dissipation in upstream direction.

The string stability transfer function thus equals $\Gamma_i(j\omega) = Z_i(s)/Z_{i-1}(s)$, with $Z(s)$ being the Laplace transform of $z(t)$. In general, however, $Z_{i-1}(s)$ is not an independent input, since it is determined by other downstream vehicles or even, in case of bidirectional communication, by upstream vehicles in the string. Consequently, the string stability

transfer function should in fact be regarded as the product of transfer functions which actually have independent inputs, being the first vehicle in the string:

$$\Gamma_i(s) = \frac{Z_i(s)}{Z_1(s)} \left(\frac{Z_{i-1}(s)}{Z_1(s)} \right)^{-1}. \quad (5.19)$$

$\Gamma_i(s)$ is capable of describing the simplest one-vehicle look-ahead communication structure, but also more complex communication structures as shown in  Fig. 5.6. In the latter case, however, $\Gamma_i(s)$ not only includes the dynamics of two neighboring vehicles, but also the dynamics of vehicles further away.

4.2.4 The z-Domain Approach

Within the framework of analysis of string stability of infinite-length vehicle strings, or any other system consisting of identical interconnected subsystems, the model of such a system is formulated in state-space as

$$\dot{\mathbf{x}}_l(t) = \sum_{j=-\infty}^{\infty} (\mathbf{A}_{l-j} \mathbf{x}_j(t) + \mathbf{B}_{l-j} \mathbf{u}_j(t)) \quad (5.20a)$$


$$\mathbf{y}_l(t) = \sum_{j=-\infty}^{\infty} \mathbf{C}_{l-j} \mathbf{x}_j(t) \quad (5.20b)$$

where $\mathbf{x}_l(t)$ denotes the state of subsystem $l \in \mathbb{Z}$ and $\mathbf{u}_j(t)$ the input of subsystem j . The matrices \mathbf{A}_{l-j} , \mathbf{B}_{l-j} and \mathbf{C}_{l-j} are the state, input, and output matrix, respectively, regarding the influence of subsystem j on subsystem l . In general, the influence of other subsystems decreases when they are “further away,” meaning that the state-space matrices approach zero for $j \rightarrow \pm\infty$. Moreover, a distributed linear output-feedback controller is assumed, according to

$$\mathbf{u}_j(t) = \mathbf{K} \mathbf{y}_j(t). \quad (5.21)$$

In order to analyze this system, it can be transformed using the bilateral \mathcal{Z} -transform. Consider to this end a sequence $\{a_k(t)\}_{k=-\infty}^{\infty}$. The bilateral \mathcal{Z} -transform $\mathcal{Z}(a_k(t)) = \hat{a}(z, t)$ is then defined by

$$\hat{a}(z, t) \triangleq \sum_{k=-\infty}^{\infty} a_k(t) z^{-k} \quad (5.22)$$

where z is a complex variable. The model of  Eq. 5.20 is already formulated as a convolution, which makes the application of the \mathcal{Z} -transform particularly easy, resulting in

$$\hat{\mathbf{x}}(z, t) = \hat{\mathbf{A}}(z) \hat{\mathbf{x}}(z, t) + \hat{\mathbf{B}}(z) \hat{\mathbf{u}}(z, t) \quad (5.23a)$$

$$\hat{\mathbf{y}}(z, t) = \hat{\mathbf{C}}(z) \hat{\mathbf{x}}(z, t) \quad (5.23b)$$

$$\hat{\mathbf{u}}(z, t) = \mathbf{K} \hat{\mathbf{y}}(z, t). \quad (5.23c)$$

Defining $\mathbf{D}(z) = \hat{\mathbf{A}}(z) + \hat{\mathbf{B}}(z)\hat{\mathbf{K}}(z)\hat{\mathbf{C}}(z)$, the closed-loop system thus reads

$$\hat{\mathbf{x}}(z, t) = \mathbf{D}(z)\hat{\mathbf{x}}(z, t). \quad (5.24)$$

Using this approach, a third type of string stability definition can be formulated (Chu 1974; El-Sayed and Krishnaprasad 1981; Barbieri 1993).

Definition 4.3

(Interconnected System String Stability). Assume a dynamical interconnected system described by an infinite number of identical n^{th} -order subsystems (Eq. 5.20) with the feedback control law of \blacktriangleright Eq. 5.21, such that the bilateral \mathcal{Z} -transform of the closed-loop system is given by \blacktriangleright Eq. 5.24. Then this system is string stable if all the eigenvalues $\lambda_i (i = 1, \dots, n)$ of $\mathbf{D}(z)$ are in the left-half complex plane, for all z on the unit circle, i.e.,

$$\text{Re}(\lambda_i(\hat{\mathbf{D}}(e^{j\theta}))) \leq 0$$

for $i = 1, \dots, n$ and for $0 \leq \theta < 2\pi$.

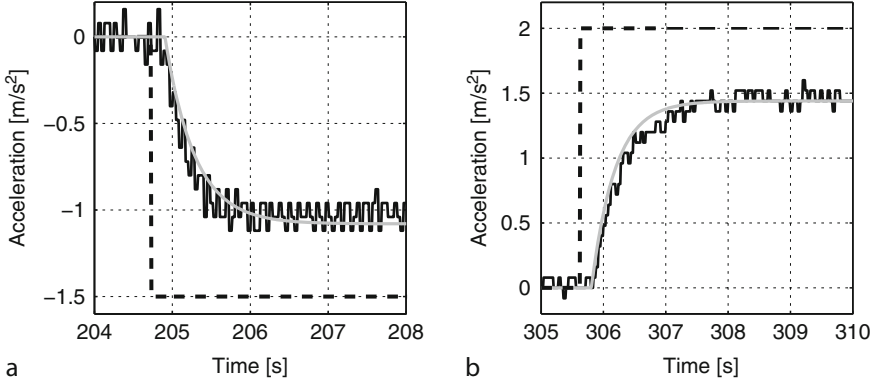
This approach, although elegant in itself, has severe limitations due to the assumed infinite length of the interconnected system and due to the fact that it only covers homogeneous strings, i.e., all subsystems must be identical.

4.2.5 Concluding Remarks on String Stability

It can probably be stated that Definition 4.1 (or generalized versions thereof) regarding \mathcal{L}_p -string stability formulates a true string stability definition. In itself, however, this definition provides little support for controller synthesis, as opposed to Definition 4.2. The latter has the character of a condition for string stability rather than a definition of this notion. It should therefore be possible to rigorously derive this condition using for instance the notion of input–output stability (Khalil 2000). As far as the string stability of interconnected systems given in Definition 4.3 is concerned, as already mentioned, this approach mainly has theoretical value regarding the stability analysis and controller synthesis of infinitely long strings of identical subsystems. In practice, the performance-oriented approach is often adopted, which is nicely illustrated in the case study, shortly summarized in the next section.

4.3 Case Study

To validate the theory, especially with respect to the performance-oriented string stability approach, experiments have been performed with two vehicles, as described in Naus et al. (2010b,c). Vehicle 1 communicates its actual acceleration to vehicle 2. The latter is equipped with an electrohydraulic braking (EHB) system, facilitating brake-by-wire control and an electronically controlled throttle valve; both serve as actuators for the custom-made lower-level acceleration controller. Finally, vehicle 2 is equipped with a laser radar.



■ Fig. 5.7

Validation step-response results for (a) braking and (b) accelerating. The reference step input (dashed black), the measurement results (solid black), and the corresponding simulation results (solid gray) are shown

The vehicle model of ● Eq. 5.6 is adopted, which is supposed to also include the lower-level acceleration controller. After experiments, the following parameter values have been identified: $k_i = 0.72$, $\tau_i = 0.38$ s, and $\theta_i = 0.18$ s. The model is validated using step responses, depicted in ● Fig. 5.7. From this figure, it can be concluded that the model of ● Eq. 5.6 adequately describes the relevant dynamics.

The controller K (refer to ● Fig. 5.5) is chosen as a lead-lag filter, according to

$$K_i(s) = \frac{\omega_{K,i}}{k_i} \frac{\omega_{K,i} + s}{\omega_{f,i} + s}, \quad \text{for } i > 1 \quad (5.25)$$

with $\omega_{K,i} = 0.5$ rad/s and $\omega_{f,i} = 100.0\pi$ rad/s. These parameters are mainly based on ensuring a sufficient closed-loop bandwidth and phase margin of the individually controlled vehicle, characterized by the complementary sensitivity $T_i(s) = (1 + G_i(s)K_i(s))^{-1} G_i(s)K_i(s)$ (see ● Sect. 4.2.1). The identified vehicle gain k_i is compensated for by including this gain in the controller as well. As mentioned, the preceding vehicle acceleration is communicated, which then serves as the input for the feedforward filter F . Hence, the feedforward filter should compensate for the vehicle dynamics $G_i(s)$ and take the spacing policy into account as well. Since it cannot compensate for time delay, the following filter is implemented:

$$F_i(s) = (H_i(s)G_i(s)s^2)^{-1}, \quad \text{for } i > 1 \quad (5.26)$$

using the constant time headway spacing policy given by ● Eq. 5.13.

For the wireless intervehicle communication, the standard WiFi protocol IEEE 802.11 g is used, with an update rate of 10 Hz. The acceleration of vehicle 1 is derived from the built-in antilock braking system (ABS), which is available on the CAN-bus, and communicated to vehicle 2. A zero-order-hold approach is adopted for

the communicated signal, introducing a corresponding average delay of about 50 ms. To synchronize the measurements of both vehicles, GPS time stamping is adopted. Correspondingly, an additional communication delay of about 10 ms is identified. Combination of these values yields $\psi_i \approx 60$ ms as a total delay. Hence, the communication delay model D_i , shown in [Fig. 5.5](#), is defined by

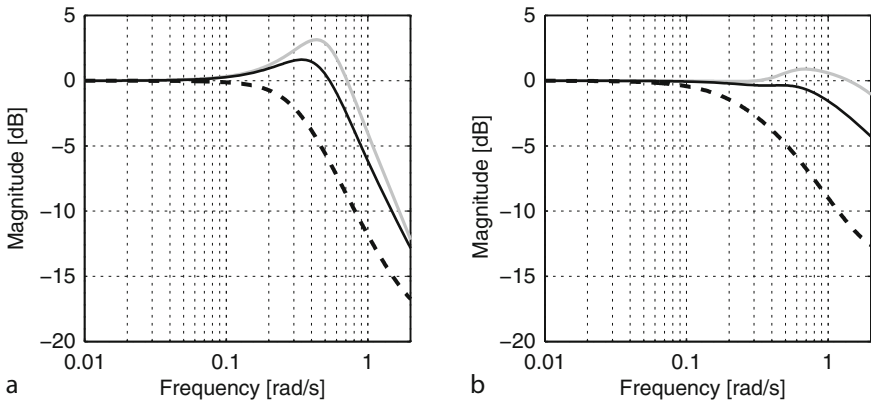
$$D_i(s) = e^{-\psi_i s}. \quad (5.27)$$

Having determined all control system components, the string stability transfer function $\Gamma_i(s)$ can be analyzed. In Naus et al. (2010b, c), it is shown that for a homogeneous vehicle string, i.e., a string with all identical subsystems, the string stability transfer functions are identical, regardless of whether the distance error, the acceleration, or the control input is chosen as input/output. The resulting string stability transfer function reads

$$\begin{aligned} \Gamma_i(s) &= \frac{(F_i(s)D_i(s)s^2 + K_i(s))G_i(s)}{1 + H_i(s)G_i(s)K_i(s)} \\ &= \frac{D_i(s) + H_i(s)G_i(s)K_i(s)}{H_i(s)(1 + H_i(s)G_i(s)K_i(s))} \end{aligned} \quad (5.28)$$

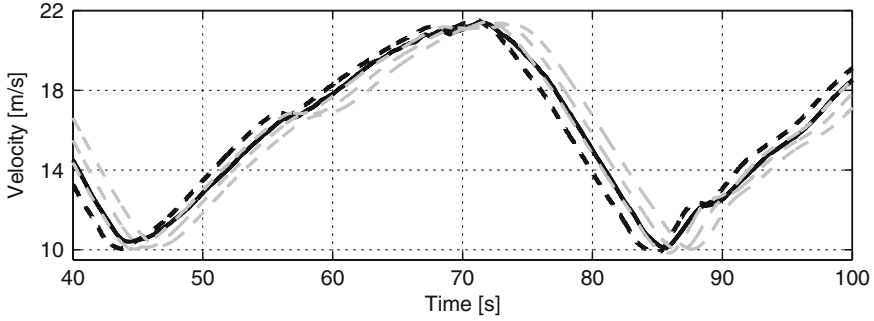
which reduces to $\Gamma_i(s) = 1/H_i(s)$ in case there is no communication delay, i.e., $D_i(s) = 1$. [Figure 5.8](#) shows the resulting magnitude for various values of the time headway h_i . Also the ACC equivalent is depicted, which in fact equals the CACC with $F_i(s) = 0$. It can be clearly seen that ACC provides string-stable behavior only for $h_i = 3.0$ s, which is considered quite large in practice. CACC provides string-stable behavior for values as small as $h_i = 1.0$ s, given the current parameters of the controller $K_i(s)$; Naus et al. (2010c) mention $h_i = 2.6$ s as minimum string-stable value for ACC and $h_i = 0.8$ s for CACC.

[Figure 5.9](#) contains some experiment results, showing the measured velocity of the first and second vehicle, extended with simulated results of three more follower vehicles



■ Fig. 5.8

Bode magnitude plots of $\Gamma_i(j\omega)$, in the case of (a) ACC, and (b) CACC, for time headway $h_i = 0.5$ s (solid gray), $h_i = 1.0$ s (solid black), and $h_i = 3.0$ s (dashed black)



■ Fig. 5.9

Measured velocity of vehicle 1 (*dashed black*) and vehicle 2 (*solid black*), as well as the resulting velocities for three simulated follower vehicles (*dashed gray*)

for CACC with a time headway $h_i = 1.0$ s. From this figure, it can be clearly seen that the disturbance, introduced by the first vehicle, is attenuated, albeit barely.

5 Roadside and Area Control

In this section, we summarize the approach proposed in Baskar et al. (2008, 2009a, b) and Baskar (2009) for the roadside controllers to determine optimal speeds, lane allocations, and on-ramp release times for the platoons, and for the area controllers to determine optimal flows and speeds on links. As control approach we adopt the model predictive control (MPC) scheme presented in ● Sect. 2.2.2.

5.1 MPC for Roadside Controllers

In order to make the MPC approach tractable, the roadside controllers do not consider each individual vehicle in each platoon separately, but they consider each platoon as a single unit and they monitor the movements of the platoons in the highway stretch under their control. More specifically, the platoons are modeled using the so-called big car model, i.e., as a single (long) vehicle with a speed-dependent length:

$$L_{\text{plat},p}(k) = (n_p - 1)S_0 + \sum_{i=1}^{n_p-1} T_{\text{head},i} v_{\text{plat},p}(k) + \sum_{i=1}^{n_p} L_i,$$

where $L_{\text{plat},p}(k)$ is the length of platoon p at time step k , n_p is the number of vehicles in the platoon, S_0 the minimum safe distance that is to be maintained at zero speed, $T_{\text{head},i}$ is the desired time headway for vehicle i in the platoon, $v_{\text{plat},p}(k)$ is the speed of the platoon (leader), and L_i is the length of vehicle i . In this way a good trade-off is obtained between computational speed and accuracy.

The control inputs determined for each platoon are its speed, lane assignment, size, as well as release time (at on-ramps) and route choice (at bifurcations). The objective function and constraints can correspond to general traffic performance criteria such as total time spent, throughput, emissions, etc., or they could reflect tracking of targets set forth by the area controllers.

In general, this results in a mixed-integer nonlinear optimization problem (if lane allocation and/or size are included in the MPC optimization) or in a real-valued nonlinear optimization problem (if lane allocation and size are assigned using heuristics or logic rules). Mixed-integer optimization problems could be solved using genetic algorithms, simulated annealing, or branch-and-bound methods. Continuous optimization problems can be solved using multi-start sequential quadratic programming, genetic algorithms, simulated annealing, or pattern search.

5.2 MPC for Area Controllers

In principle, the optimal route choice control problem in IVHS consists in assigning an optimal route to each individual platoon in the network. However, this results in a huge nonlinear integer optimization problem with high computational complexity and requirements, making the problem in fact intractable in practice. Since considering each individual platoon is too computationally intensive for online real-time control, the area controllers consider a more aggregate model based on flows of platoons. In this context two approaches have been pursued, namely, one based on a flow-and-queue model (Baskar 2009; Baskar et al. 2009a) and one based on a METANET-like model for platoons in an IVHS (Baskar 2009; Baskar et al. 2009b).

In the first approach, the evolution of the flows (on highway stretches) and queue lengths (at junctions) in the network is described using simple queuing models and assuming a fixed average speed in each highway stretch. The control decisions are then the assignment of flows to the links. Although in general this results in a nonlinear, non-convex, and nonsmooth optimization problem, it was shown in Baskar (2009) and Baskar et al. (2009a) that the resulting optimization problem can be approximated using mixed-integer linear programming (MILP), for which efficient branch-and-bound solvers are currently available (Fletcher and Leyffer 1998). The MILP solution can then be applied directly to the IVHS or it can be used as a good initial starting point for a local optimization of the original nonlinear, non-convex optimization problem.

The second approach is based on a reformulation of the macroscopic traffic flow model METANET (Messmer and Papageorgiou 1990; Kotsialos et al. 2002) for IVHS. The resulting IVHS–METANET model describes the evolution of the traffic flows through average densities, flows, and speeds in the highway segments. The control decisions in this case are the splitting rates at the network nodes and possibly also the speeds on the links. This then results in a nonlinear, non-convex optimization problem with real-valued variables. To solve the nonlinear optimization problem, we can use a global or

a multi-start local optimization method such as multi-start sequential quadratic programming, pattern search, genetic algorithms, or simulated annealing.

5.3 Interfaces Between the Different Control Layers

The higher-level controllers can influence the controller in the level immediately below them in two ways: by specifying weights, set-points, or reference signals in the objective function, or by specifying targets or thresholds in the constraints. The lower-level controller then has to solve an optimization problem of the form

$$\min_{\mathbf{u}(k), \dots, \mathbf{u}(k+N_p-1)} J(k) = J_{\text{high}}(k) + \lambda J_{\text{local}}(k) \quad (5.29)$$

$$\text{s.t. } \mathbf{x}(k+j+1) = \mathbf{f}(\mathbf{x}(k+j), \mathbf{u}(k+j), \mathbf{d}(k+j)) \text{ for } j = 0, \dots, N_p - 1 \quad (5.30)$$

$$\mathbf{u}(k+j) = \mathbf{u}(k+N_c-1) \text{ for } j = N_c, \dots, N_p - 1 \quad (5.31)$$

$$\mathbf{C}_{\text{high}}(\mathbf{x}(k), \dots, \mathbf{x}(k+N_p), \mathbf{u}(k), \dots, \mathbf{u}(k+N_c-1)) \quad (5.32)$$

$$\mathbf{C}_{\text{local}}(\mathbf{x}(k), \dots, \mathbf{x}(k+N_p), \mathbf{u}(k), \dots, \mathbf{u}(k+N_c-1)) \quad (5.33)$$

where J_{high} and \mathbf{C}_{high} represent, respectively, the objectives and constraints (in the form of a system of equations and/or inequalities) imposed by the higher-level controller, J_{local} is the local, additional objective that has to be optimized, $\lambda > 0$ is a weighting factor, $\mathbf{C}_{\text{local}}$ contains the local constraints, and $\mathbf{x}(k+j)$ is the prediction of the state of the traffic system (region, area, highway stretch, depending on the control level) at time step $k+j$, $\mathbf{u}(k+j)$ is the control input at time step $k+j$, and $\mathbf{d}(k+j)$ is the estimate of the traffic demand at time step $k+j$. In addition, the model equations (◆ Eq. 5.30) and the control horizon constraint ◆ Eq. 5.31 are also included.

The control variables determined by the area controllers are the flows on the links and/or the splitting rates at the nodes with more than one outgoing link (and if speed limits are included, also these speed limits). Once the optimal flows, splitting rates, and speeds have been determined by the area controller, they are sent to the lower-level roadside controllers, which can then translate them into actual speed, route, and lane allocation instructions for the platoons. So in this case the communication goes through the performance criterion J_{high} .

The roadside controllers can provide lane allocation commands and speeds in order to realize the target flows and speeds in the links. These control measures can then slow down or speed up platoons in the links and also steer the platoons in certain directions depending on the imposed splitting rates for the flows. At the nodes, the roadside controller will additionally provide routing instructions for every platoon on the stretch under its supervision. The roadside controller will determine these routing instructions by taking into account the destinations of the platoons and also the imposed splitting rates or the target flows on the adjacent highways.

The roadside controller can combine the speed and route guidance control measures along with on-ramp access timing to control the platoons that enter from on-ramps. The platoon length will play a crucial role while providing routing instructions to the platoons at internal nodes or bifurcation junctions. So if necessary, the roadside controllers can then also provide commands for platoon splits and merges, and determine new platoon compositions and platoon lengths.

6 Challenges and Open Issues

Now we discuss the main technological, economical, and societal challenges that will have to be addressed when actually implementing an IVHS system.

Although several authors have indicated how control design methods such as static feedback control, MPC, and AI-based control could be used for IVHS and traffic management and control systems based on intelligent vehicles and platoons (see the preceding sections), real integration of these methods is still lacking. This is one of the challenges that still have to be addressed.

Moreover, since IVHS and IVs are nonlinear and often even hybrid (i.e., they exhibit both continuous dynamics and discrete-event behavior [switching]), properties such as stability and robustness of the traffic system have also to be investigated further. In addition, in particular at the platoon level and higher, there is also a need for performance guarantees supported by solid fundamental results.

Two other important remaining open problems in control of IVHS are platoon formation and control, and scalability.

In literature there are no strict rules available on how to form platoons and on how many vehicles to include in a platoon. This can either be specific to a given road or to a destination. There are few articles that deal with vehicle sorting with respect to platoon sizes and platoon formation time, and also on the design of platoon maneuver protocols (Hsu et al. 1991; Hall and Chin 2005). Moreover, the design of platoon controllers is certainly not standardized. Especially the particular choice of spacing policy will heavily influence the overall IVHS performance with respect to throughput, safety, and fuel consumption. Although some research has been done, this is still an open issue.

Some of IVHS frameworks such as the PATH, Auto21 CDS, and Dolphin frameworks are by nature hierarchical and offer thus a certain degree of scalability with regard to network size. Other frameworks such as PReVENT, SafeSpot, and CVIS are not explicitly hierarchical and are thus not inherently scalable with regard to network size. However, none of the frameworks explicitly addresses scalability and the scalability of the frameworks has not yet been investigated in detail in literature. So this is also a topic for future research.

The technical issues outlined above are still open and need to be addressed. Moreover, the IVHS approach requires major investments to be made by both the government (or the body that manages the highway system) and the constructors and owners of the vehicles. Since few decisions are left to the driver and since the AHS assumes almost

complete control over the vehicles, which drive at high speeds and at short distances from each other, a strong psychological resistance to this traffic congestion policy is to be expected. In addition, the fact that vehicles can be tracked through the entire network may raise some concerns regarding privacy and liability issues.

Another important question is how the transition of the current highway system to an AHS-based system should occur and – once it has been installed – what has to be done with vehicles that are not yet equipped for IVHS. Some of the transition issues that have to be taken into account are (Fenton 1994): How will the system be funded? What types of accidents can be expected to occur in AHS, in what numbers, and with what consequences? Will bidirectional communication and transfer of information be allowed between IVs and roadside infrastructure? What are the legal implications of an accident, especially if it were caused by system error or a system oversight? How will an AHS implementation be coordinated on an international level?

7 Conclusions

In this chapter we have presented an overview of traffic management and control frameworks for IVHS. First, we have given a short survey of the main control design methods that could be used in IVHS. Next, we have discussed various traffic management architectures for IVHS such as PATH, Dolphin, Auto21 CDS, CVIS, SafeSpot, and PReVENT. Subsequently, we have focused in more detail on the platoon, roadside, and area layers. Finally, we have identified some open issues and future challenges in the further implementation and actual deployment of IVHS traffic management systems, in particular, integration, stability, scalability, and transition issues.

References

- Aamodt A, Plaza E (1994) Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun* 7(1):39–59
- Åström KJ, Wittenmark B (1997) Computer-controlled systems – theory and applications, 3rd edn. Prentice-Hall, Upper Saddle River
- Barbieri E (1993) Stability analysis of a class of interconnected systems. *ASME J Dyn Syst Meas Control* 115(3):546–551
- Baskar LD (2009) Traffic management and control in intelligent vehicle highway systems. Ph.D. thesis, Delft University of Technology, Delft, TRAIL Thesis Series T2009/12
- Baskar LD, De Schutter B, Hellendoorn H (2007) Hierarchical traffic control and management with intelligent vehicles. In: Proceedings of the 2007 IEEE intelligent vehicles symposium (IV'07), Istanbul, pp 834–839
- Baskar LD, De Schutter B, Hellendoorn J (2008) Model-based predictive traffic control for intelligent vehicles: dynamic speed limits and dynamic lane allocation. In: Proceedings of the 2008 IEEE intelligent vehicles symposium (IV'08), Eindhoven, pp 174–179
- Baskar LD, De Schutter B, Hellendoorn H (2009a) Optimal routing for intelligent vehicle highway systems using mixed integer linear programming. In: Proceedings of the 12th IFAC symposium on transportation systems, Redondo Beach, pp 569–575
- Baskar LD, De Schutter B, Hellendoorn J (2009b) Optimal routing for intelligent vehicle highway

- systems using a macroscopic traffic flow model. In: Proceedings of the 12th international IEEE conference on intelligent transportation systems (ITSC 2009), St. Louis, pp 576–581
- Baskar LD, De Schutter B, Papp Z, Hellendoorn J (2011) Traffic control and intelligent vehicle highway systems: a survey. *IET Intell Transp Syst* (to appear)
- Bellman R (1957) Dynamic programming. Princeton University Press, Princeton
- Bishop R (2005) Intelligent vehicles technology and trends. Artech House, Norwood
- Broucke M, Varaiya P (1997) The automated highway system: a transportation technology for the 21st century. *Control Eng Pract* 5(11):1583–1590
- Chen SH, Jakeman AJ, Norton JP (2008) Artificial intelligence techniques: an introduction to their use for modelling environmental systems. *Math Comput Simul* 78(2):379–400
- Chu K-C (1974) Optimal decentralized regulation for a string of coupled systems. *IEEE Trans Autom Control* 19(3):243–246
- CVIS web site. <http://www.cvisproject.org/>. Accessed 15 Nov 2010
- Daganzo CF (1997) Fundamentals of transportation and traffic operations. Pergamon, Oxford
- Desoer CA, Vidyasagar M (2009) Feedback systems: input-output properties classics in applied mathematics. Society for Industrial and Applied Mathematics (SIAM), Pennsylvania
- El-Sayed ML, Krishnaprasad PS (1981) Homogeneous interconnected systems: an example. *IEEE Trans Autom Control* 26(4):894–901
- Fenton RE (1994) IVHS/AHS: driving into the future. *IEEE Control Syst Mag* 14(6):13–20
- Ferber J (1999) Multi-agent systems – an introduction to distributed artificial intelligence. Addison-Wesley, Harlow
- Fletcher R, Leyffer S (1998) Numerical experience with lower bounds for MIQP branch-and-bound. *SIAM J Optim* 8(2):604–616
- Gehring O, Fritz H (1997) Practical results of a longitudinal control concept for truck platooning with vehicle to vehicle communication. In: Proceedings of the IEEE conference on intelligent transportation systems, Boston, pp 117–122
- Gelfand M, Fomin SV (1991) Calculus of variations. Dover, New York
- Hall R, Chin C (2005) Vehicle sorting for platoon formation: impacts on highway entry and throughput. *Transp Res C* 13(5–6):405–420
- Hallé S, Chaib-draa B (2005) A collaborative driving system based on multiagent modelling and simulations. *Transp Res C Emerg Technol* 13(4):320–345
- Hammerstrom D (1993) Working with neural networks. *IEEE Spectr* 30(7):46–53
- Hayes-Roth F (1985) Rule-based systems. *Commun ACM* 28(9):921–932
- Hedrick JK, Tomizuka M, Varaiya P (1994) Control issues in automated highway systems. *IEEE Control Syst Mag* 14(6):21–32
- Hestenes MR (1966) Calculus of variations and optimal control theory. Wiley, New York
- Horowitz R, Varaiya P (2000) Control design of an automated highway system. *Proc IEEE* 88(7):913–925
- Hsu A, Eskafi F, Sachs S, Varaiya P (1991) Design of platoon maneuver protocols for IVHS. Technical Report 96–21, California Partners for Advanced Transit and Highways PATH, University of California, Berkeley
- Hsu A, Eskafi F, Sachs S, Varaiya P (1993) Protocol design for an automated highway system. *Discrete Event Dyn Syst Theory Appl* 2(1):183–206
- International Organization for Standardization (2002) Transport information and control systems – adaptive cruise control systems – performance requirements and test procedures, ISO 15622. Technical report, Transport information and control systems
- Ioannou PA, Chien CC (1993) Autonomous intelligent cruise control. *IEEE Trans Vehicle Technol* 42(4):657–672
- Kachroo P, Özbay K (1999) Feedback control theory for dynamic traffic assignment. *Advances in industrial control*. Springer, Berlin
- Khalil HK (2000) Nonlinear systems, 3rd edn. Prentice-Hall, Upper Saddle River
- Kirk DE (1970) Optimal control theory: an introduction. Prentice-Hall, Englewood Cliffs
- Klinge S, Middleton RH (2009) Time headway requirements for string stability of homogeneous linear unidirectionally connected systems. In: Proceedings of the 48th IEEE conference on decision and control, Shanghai, pp 1992–1997
- Klir GJ, Yuan B (1995) Fuzzy sets and fuzzy logic: theory and applications. Prentice-Hall, Upper Saddle River
- Kotsialos A, Papageorgiou M, Diakaki C, Pavlis Y, Middelham F (2002) Traffic flow modeling of

- large-scale motorway networks using the macroscopic modeling tool METANET. *IEEE Trans Intell Transp Syst* 3(4):282–292
- Levine WS, Athans M (1966) On the optimal error regulation of a string of moving vehicles. *IEEE Trans Autom Control* 11(3):355–361
- Li K, Ioannou P (2004) Modeling of traffic flow of automated vehicles. *IEEE Trans Intell Transp Syst* 5(2):99–113
- Maciejowski JM (2002) Predictive control with constraints. Prentice-Hall, Harlow
- Messmer A, Papageorgiou M (1990) METANET: a macroscopic simulation program for motorway networks. *Traffic Eng Control* 31(9):466–470
- Naus GJL, Vugts R, Ploeg J, van de Molengraft MJG, Steinbuch M (2009) Toward on-the-road implementation of cooperative adaptive cruise control. In: *Proceedings of the 16th world congress & exhibition on intelligent transport systems and services*, Stockholm
- Naus GJL, Ploeg J, van de Molengraft MJG, Heemels WPMH, Steinbuch M (2010a) Design and implementation of parameterized adaptive cruise control: an explicit model predictive control approach. *Control Eng Pract* 18(8):882–892
- Naus GJL, Vugts R, Ploeg J, van de Molengraft MJG, Steinbuch M (2010b) Cooperative adaptive cruise control, design and experiments. In: *Proceedings of the American control conference*, Baltimore, pp 6145–6150
- Naus GJL, Vugts R, Ploeg J, van de Molengraft MJG, Steinbuch M (2010c) String-stable CACC design and experimental validation: a frequency-domain approach. *IEEE Trans Vehicle Technol* 59(9):4268–4279
- Nguyen HT, Walker EA (1999) *A first course in fuzzy logic*, 2nd edn. Chapman & Hall, Boca Raton
- Papageorgiou M (1983) Applications of automatic control concepts to traffic flow modeling and control. *Lecture Notes in Control and Information Sciences*. Springer, Berlin, Germany
- Pardalos PM, Resende MGC (2002) *Handbook of applied optimization*. Oxford University Press, Oxford
- PREVENT web site. <http://www.prevent-ip.org/>. Accessed 15 Nov 2010
- Rajamani R (2006) *Vehicle dynamics and control*. Mechanical engineering series. Springer, New York
- Rao BSY, Varaiya P (1994) Roadside intelligence for flow control in an IVHS. *Transp Res C* 2(1):49–72
- Rawlings JB, Mayne DQ (2009) *Model predictive control: theory and design*. Nob Hill, Madison
- Ritchie SG (1990) A knowledge-based decision support architecture for advanced traffic management. *Transp Res A* 24(1):27–37
- Russell S, Norvig P (2003) *Artificial intelligence: a modern approach*. Prentice-Hall, Englewood Cliffs
- SafeSpot web site. <http://www.safespot-eu.org/>. Accessed 15 Nov 2010
- Sheikholeslam S, Desoer CA (1992) A system level study of the longitudinal control of a platoon of vehicles. *ASME J Dyn Syst Meas Control* 114(2):286–292
- Sheikholeslam S, Desoer CA (1993) Longitudinal control of a platoon of vehicles with no communication of lead vehicle information: a system level study. *IEEE Trans Vehicle Technol* 42(4):546–554
- Shladover SE (2007) PATH at 20 – history and major milestones. *IEEE Trans Intell Transp Syst* 8(4):584–592
- Stanković SS, Stanojević MJ, Šiljak DD (2000) Decentralized overlapping control of a platoon of vehicles. *IEEE Trans Control Syst Technol* 8(5):816–832
- Sussman JM (1993) Intelligent vehicle highway systems: challenge for the future. *IEEE Micro* 1(14–18):101–104
- Sussmann HJ, Willems JC (1997) 300 years of optimal control: from the brachistochrone to the maximum principle. *IEEE Control Syst Mag* 17(3):32–44
- Sutton RS, Barto AG (1998) *Reinforcement learning: an introduction*. MIT Press, Cambridge
- Swaroop D, Hedrick JK (1996) String stability of interconnected systems. *IEEE Trans Autom Control* 41(3):349–357
- Swaroop D, Rajagopal KR (1999) Intelligent cruise control systems and traffic flow stability. *Transp Res C Emerg Technol* 7(6):329–352
- Swaroop D, Hedrick JK, Chien CC, Ioannou P (1994) A comparison of spacing and headway control laws for automatically controlled vehicles. *Veh Syst Dyn* 23(1):597–625
- Swaroop D, Hedrick JK, Choi SB (2001) Direct adaptive longitudinal control of vehicle platoons. *IEEE Trans Vehicle Technol* 50(1):150–161
- Toulminet G, Boussuge J, Laureau C (2008) Comparative synthesis of the 3 main European projects dealing with cooperative systems (CVIS,

- SAFESPOT and COOPERS) and description of COOPERS demonstration site 4. In: Proceedings of the 11th IEEE conference on intelligent transportation systems, Beijing, pp 809–814
- Tsugawa S, Kato S, Tokuda K, Matsui T, Fujii H (2000) An architecture for cooperative driving of automated vehicles. In: Proceedings of the IEEE intelligent transportation symposium, Dearborn, pp 422–427
- Tsugawa S, Kato S, Tokuda K, Matsui T, Fujii H (2001) A cooperative driving system with automated vehicles and inter-vehicle communications in demo 2000. In: Proceedings of the IEEE conference on intelligent transportation systems, Oakland, pp 918–923
- Varaiya P (1993) Smart cars on smart roads: problems of control. *IEEE Trans Autom Control* 38(2): 195–207
- Varaiya P, Shladover SE (1991) Sketch of an IVHS systems architecture. In: Vehicle navigation and information systems, Dearborn, pp 909–922
- Venhovens P, Naab K, Adiprasito B (2000) Stop and go cruise control. *Int J Automot Technol* 1(2): 61–69
- Weiss G (ed) (1999) Multiagent systems: a modern approach to distributed artificial intelligence. MIT Press, Cambridge
- Yao X (1999) Evolving artificial neural networks. *Proc IEEE* 87(9):1423–1447

6 Behavioral Adaptation and Acceptance

Marieke H. Martens^{1,2} · Gunnar D. Jenssen³

¹TNO, Soesterberg, The Netherlands

²University of Twente, Enschede, The Netherlands

³Transport Research, SINTEF, Trondheim, Norway

1	<i>Introduction</i>	118
1.1	Direct Behavioral Adaptation	119
1.2	Indirect Behavioral Adaptation	119
2	<i>Models of Behavioral Adaptation</i>	120
3	<i>Types of Behavioral Adaptation</i>	123
4	<i>When Does BA Occur?</i>	125
5	<i>Behavioral Adaptation in a Longer-Term Perspective</i>	125
6	<i>Relation Behavioral Adaptation and Safety</i>	128
7	<i>Behavioral Adaptation and Acceptance</i>	130
8	<i>Future Directions</i>	133
9	<i>Will New Skills Appear?</i>	135
10	<i>International Aspects of Adaptation to ADAS</i>	135
11	<i>General Conclusions</i>	136

Abstract: One purpose of Intelligent Vehicles is to improve road safety, throughput, and emissions. However, the predicted effects are not always as large as aimed for. Part of this is due to indirect behavioral changes of drivers, also called behavioral adaptation. Behavioral adaptation (BA) refers to unintended behavior that arises following a change to the road traffic system. Qualitative models of behavioral adaptation (formerly known as risk compensation) describe BA by the change in the subjectively perceived enhancement of the safety margins. If a driver thinks that the system is able to enhance safety and also perceives the change in behavior as advantageous, adaptation occurs. The amount of adaptation is (indirectly) influenced by the driver personality and trust in the system. This also means that the amount of adaptation differs between user groups and even within one driver or changes over time.

Examples of behavioral change are the generation of extra mobility (e.g., taking the car instead of the train), road use by “less qualified” drivers (e.g., novice drivers), driving under more difficult conditions (e.g., driving on slippery roads), or a change in distance to the vehicle ahead (e.g., driving closer to a lead vehicle with ABS).

In effect predictions, behavioral adaptation should be taken into account. Even though it may reduce beneficial effects, BA (normally) does not eliminate the positive effects. How much the effects are reduced depends on the type of ADAS, the design of the ADAS, the driver, the current state of the driver, and the local traffic and weather conditions.

1 Introduction

The introduction of ITS (Intelligent Transportation Systems) is generally seen as a positive step toward reducing crash risk and improving road safety. ITS includes informing, warning, and actively supporting the driver in his or her driving task by means of ADAS (Advanced Driver Assistance Systems). However, when the safety effects that result from the introduction of these safety systems are estimated, an important bias may occur. The following example illustrates this bias:

“In 30% of all accidents, driver fatigue was the major cause. The introduction of a fatigue alertness system will therefore reduce the number of accidents with 30%.”

This bias is a bias since accidents are caused by a complexity of factors, and may not be the result of fatigue only. Even more importantly, safety systems can have unintended effects on driver behavior that offset – or even negate – some of the intended benefits. What if we design systems that are not acceptable and are therefore being switched off, what if drivers continue to drive for longer periods of time since they know they will be warned, or even worse show other types of behavior in order to prevent alarms from going off since the alarm is extremely annoying? To what extent do drivers actually use and drive with the systems as intended by automotive engineers?

In terms of driver psychology, the expression “behavioral adaptation” (BA) refers to the collection of unintended behavior(s) that arises following a change to the road traffic system. Although their effect on road safety can be positive, negative, or neutral, it is the unintended and negative consequences of behavioral adaptation that are of primary

concern to road safety professionals. Wilde was one of the first to start this discussion with the introduction of the Risk Homeostasis Theory (e.g., Wilde 1982, 1988, 1994). This theory was the basis for a large number of road safety studies in the 1990s, introducing the term “risk compensation.” This describes the phenomenon that drivers adjust their behavior based on their perceived risk. Behavior may start to be less cautious in case of the introduction of safety enhancing systems (since the perceived risk decreases), and may be more cautious in case of unsafe conditions (rain, snow, darkness, not wearing seatbelts, so an increased perceived risk). However, the terms “compensation” and “homeostasis” were not considered to be adequate. Even though behavioral changes occur, there is hardly any null effect. Therefore, the term “behavioral adaptation” is used nowadays instead of “risk compensation.”

1.1 Direct Behavioral Adaptation

It is useful to note that ADAS may have direct and indirect effects on driver behavior. The direct effects on driver behavior are realized through system parameters set by the vehicle manufacturer. These direct effects are often called *engineering effects*, which are intended by system designers and implied by the systems functional specifications. For example, in an ACC (Adaptive Cruise Control) system headway distance is monitored and adjusted based on sensors with an electronic link directly to the engine, ABS, and ESC. ACC monitors the longitudinal area of safe travel in relation to stop distance parameter limits, set by the manufacturer and the driver only controls pre set choice options of the HMI. The general connotation of the concept is that it is beneficial to safety effects by changing the car following behavior. Yet, not all ADAS are specifically marketed as safety systems. For example, ACC is primarily marketed as a comfort system, although it obviously may have beneficial safety effects due to direct engineering effects on safe headway distance.

1.2 Indirect Behavioral Adaptation

It should be noted that in traffic research, behavioral adaptation most often refers specifically to the unintended and therefore indirect effects as defined by OECD (1990), “Those behaviors which may occur following the introduction of changes to the road-vehicle-user system and which were not intended by the initiators of the change.” (p. 23). The general connotation of the concept is that it is detrimental to the beneficial safety effects. However, there may also be unintentional positive safety effects of behavioral adaptation to ADAS, e.g., the increased use of turn-signal among drivers with Lane Departure Warning Systems (LDWS). In contrast to direct behavioral changes which are intended by engineers, designers, manufacturers, the OECD (1990) definition only refers to unintended effects, which hence all are indirect in nature. In other words the driver is the x factor in the equation, which may to a larger or lesser extent moderate the direct intended safety effects.

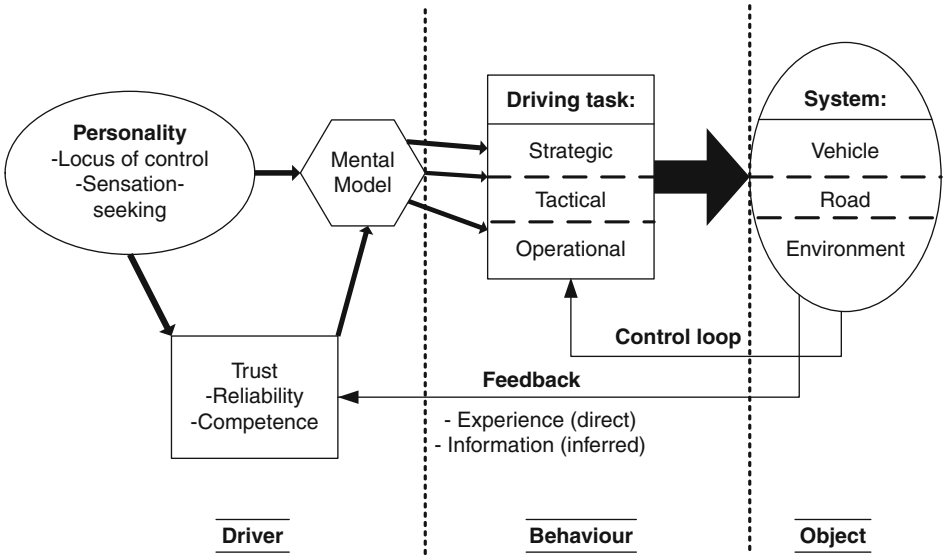
The most widely known example of behavioral adaptation to driver assistance technology is probably the case of Antilock Brake System (ABS). According to engineering

predictions of ABS' impact, it should only affect lateral control and stopping distance (OECD 1990; Vaa et al. 2006; Smiley 2000). However, studies by Fosser et al. (1997) show that drivers with ABS equipped cars drive with shorter following distances than drivers without ABS. Other well-documented examples of behavioral adaptation in road traffic is that road users have a narrowing of the eye scanning area at high speeds and in car following situations, that speed increases with increasing lane width, wider shoulders, and better road surface (OECD 1990; Mourant and Rockwell 1970; Smiley 2000). Thus, adaptation to a change is predictable and may appear in many different aspects of the driving task such as a change in headway, overtaking rate, lane change frequency, speed, braking, attention, motivation, etc. That adaptation will occur is predictable. According to Smiley (2000) we should be more surprised by its absence.

2 Models of Behavioral Adaptation

These factors are summarized in a qualitative model of Behavioral Adaptation Rudin-Brown and Noy (2002) (🔗 Fig. 6.1).

According to the model, behavioral adaptation may occur on all the levels of the driving task as defined by Michon (1985). At a strategic level, ADAS may affect the decision to drive, both in negative and positive ways. Driver monitoring systems may (indirectly) encourage a sleepy driver to keep on driving, when he or she otherwise might have stopped. ACC and collision avoidance systems may encourage drivers to keep on driving in fog or heavy rain when they otherwise would have stopped. Navigation systems



■ Fig. 6.1
Qualitative model of BA by Rudin-Brown and Noy (2002)

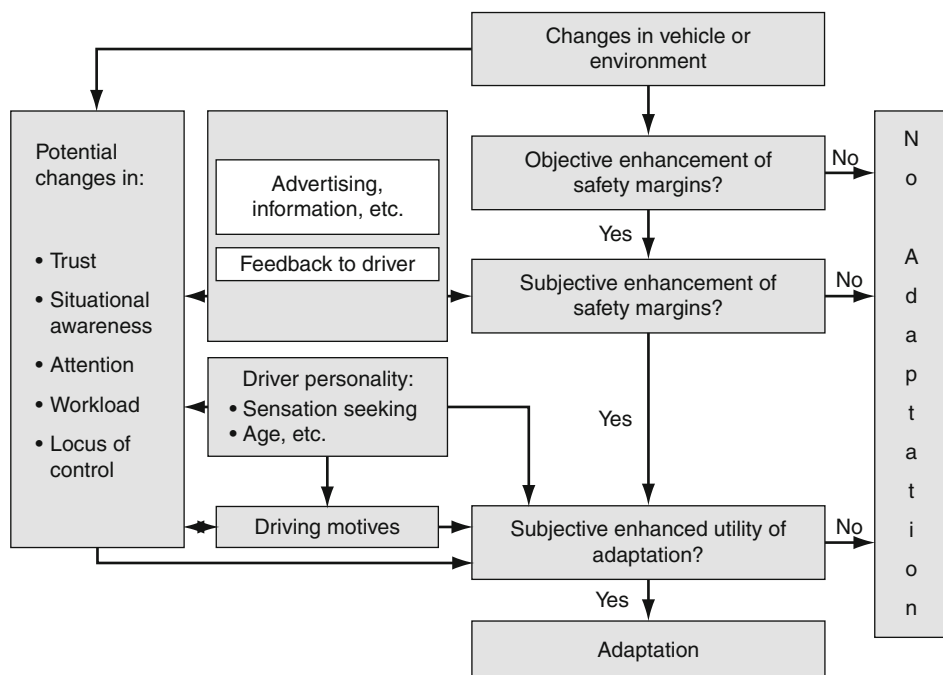
with traffic information may lead to a decision to stay home or a shift to public transport when a critical link in the road network is blocked. At a tactical level, a blind spot warning system may lead to an increased number of lane changes and overtaking maneuvers. On an operational level, increased visibility under night time driving with Adaptive Front Lighting Systems (AFS) or Vision Enhancement may lead to higher driving speeds.

The qualitative model of behavioral adaptation (Rudin-Brown and Noy 2002) gives perhaps the best account of the important components in behavioral adaptation (personality, trust, mental model), and its effect on different levels of the driving task (strategic, tactical, operational). The model does not, however, describe relevant feedback and the impact of the control loop, which may differ dependent on vehicle characteristics (e.g., ADAS, IVIS). For example when an ADAS like ACC is activated the driver is out of the loop in terms of acceleration and deceleration control actions. The driver is only in the loop if he or she monitors the process and decides to intervene (some may use the spare capacity ACC system assistance offers to send text messages, glance at incoming mail etc.) The ACC sensors take over the driver detection of headway and have a direct impact on headway distance with a feed forward loop to the traffic situation as the movement of the ACC equipped car can be observed by other road users. This feedback loop to other road users is based on characteristics of system function, not on driver actions. IVIS is on the other hand has an indirect impact on the traffic situation as observed by others. IVIS is based on driver detection of displayed system information or warnings and resultant driver actions (driver in the loop).

Similar aspects are characterized in a process model developed by Weller and Schlag (2004) (● Fig. 6.2) and named by Bjørnskau (1994, cited in Elvik and Vaa 2004). Compared to the model of Rudin-Brown and Noy, this model explicitly addresses potential changes in a Situational Awareness, attention, and workload. For example, as ADAS relieves the driver from certain driving tasks, the reduced workload may be seen as an opportunity to focus on other tasks, e.g., talking on the cell phone or reading an email on a SmartPhone.

The main focus of BA lies in the subjectively perceived enhancement of the safety margins. If a driver thinks that the system is able to enhance safety (e.g., by means of information from advertisements) and the driver also perceives the change in behavior as advantageous (subjective enhanced utility of adaptation), adaptation will occur. That is, the amount of adaptation is still (indirectly) influenced by the driver personality, his or her driving motives, and via trust, workload, locus of control, etc. This also means that the amount of adaptation differs between different user groups and may even vary within one driver (e.g., is a driver in a hurry or not?).

Many of the early models on behavioral adaptation refer to the *Peltzman Effect* which is the hypothesized tendency of people to react to a safety regulation by increasing other risky behavior, offsetting some or all of the benefit of the safety regulation (Peltzman 1975). Peltzman, an economist and true behaviorist, is however criticized for leaving out crucial human factors dimensions of response, which may be highly relevant to new advanced driver assistance systems (Smiley 2000; Carsten 2001). Examples are human factor dimensions like complacency, human error, and the well-known effects of workload, leading at one end of the spectrum to underload, low arousal, and loss of situation



■ Fig. 6.2

Process model of behavioral adaptation by Weller and Schlag (2004)

awareness, and at the other end of the spectrum to overload and stress leading to poor performance.

Under what circumstances is BA expected to be most safety critical? It is suggested that the most dangerous situation is low workload induced by driving with ADAS followed by a critical high-workload event, which could occur if a driver assistance system is not able to cope with a situation and therefore driver intervention is required (Carsten and Nilsson 2001). This is the type of situation ACC manuals now warn for, e.g., when a car suddenly changes lane and cuts off a driver, the driver may after a long drive have low workload and little attention, due to low traffic density, the monotonous task, and overreliance on the ACC system to keep a safe distance. The narrow beam of the ACC sensor is not sufficient to respond to vehicles outside the lateral system envelope. Drivers do not necessarily understand the limitations of ADAS imposed by the designer and inherent sensor limitations. This is exactly the reason why these systems are sold as comfort systems and not so much as safety systems. The driver is still responsible for safety and needs to act when the system fails. Whether drivers are always aware of this responsibility remains unknown.

Engström and Hollnagel (2005) state that BA models specifically aimed at the interaction with ADAS/IVIS functions are considerably less common than the substantial number of more generic driver behavior models. They argue particularly that a generic model of drivers' interactions with in-vehicle systems and Advanced Driver Assistance Systems, including behavioral adaptation is still lacking.

Even though there are models that include part of the elements of a generic driver behavior model including ADAS/IVIS impact, such as the COCOM and UCOM models (Hollnagel and Woods 2005), these models do not leave room for describing how aspects of an ADAS function may partially or completely take the driver out of the loop and how behavior of ADAS equipped vehicles affect surrounding traffic. Many of the older models have their basis in the 1970s, not allowing to include aspects like the introduction of ADAS. A good model has a predictive value, allowing to predict outcomes of the process in situations where the contingencies for the model change, as is the case with ADAS. ADAS like ACC and Intelligent Speed Adaptation (ISA) will change the way we drive and, consequently, will change the basis for theories, models, and tools used to explain or estimate traffic flow, capacity, safety impact, etc. For further discussion, see Modeling Driver Behavior in Automotive Environments (Cacciabue 2007).

3 Types of Behavioral Adaptation

From the models presented above, it is evident that behavioral adaptation may lead to a wide range of possible changes in driver behavior. Changes in behavior may be grouped into the following categories:

- Perceptive changes (seeing, hearing, feeling)
- Cognitive changes (comprehending, interpreting, prioritizing, selecting, deciding)
- Performance changes (driving, system handling, error)
- Driver state changes (attentiveness/awareness, workload, stress, drowsiness)
- Attitudinal changes (acceptance, rejection, overreliance, mistrust)
- Changes in adaptation to environmental conditions (weather, visibility, etc.)

Driver support systems, such as ACC, extend a driver's perceptual capabilities, since the system accounts for continuous monitoring of headway distance unaffected by fatigue. Sensors of the ACC system extend the possibility to detect lead cars in fog conditions beyond what is possible with the human eye. Future ADAS based on car-to-car communication and car-to-infrastructure systems (so-called cooperative systems) will extend the perceptual capabilities of ADAS even further. This is positive in terms of traffic safety as long as the driver does not (mis)use this benefit by driving faster than he or she otherwise would or by choosing a time headway shorter than what is safe.

There are several ways in which behavioral adaptation may influence safety. We will present different forms of behavioral adaption with some examples.

1. The generation of extra mobility. Using a driver support system can increase the amount of kilometers a person drives per year. For example, on the one hand, navigation systems reduce excess mileage because of more direct routes without people getting lost. On the other hand, they generate extra mileage into areas that were formerly avoided. Entrepreneurs who formerly "lost" 5% or 6% of the mileage driven by their fleet, because drivers selected nonoptimal routes to their destination,

may now plan an extra trip a day because navigation performance has become flawless. Also, drivers may feel more secure driving in unfamiliar areas, taking the car instead of the bus.

2. Road use by “less qualified” segments of the driving population. It is to be expected that some categories of users that did not dare to venture out in busy traffic, realizing their own imperfections, will do so if offered an extra amount of “built in” safety by means of driver support systems (e.g., elderly drivers).
3. Driving under more difficult conditions, e.g., driving at night time or on slippery roads. Due to night vision systems, the extra visual aids offered will tempt road users to drive at night, whereas they avoided these situations before. Also, having winter tires on the car drives people into harsh winter conditions, whereas they used to stay home if they would not have these tires.
4. Change in driving speed, e.g., driving faster with a new car since the brakes are more effective than the brakes on the older car that you used to drive.
5. Change in distance to the vehicle ahead, e.g., driving closer to a lead vehicle with ABS.
6. Driving less alert or concentrated, e.g., trying to read your mail on your Smartphone while driving, knowing that you are driving with a lane departure warning system that will warn you when you will drive out of your lane.
7. Avoiding, misleading, or compensating for interventions by the system, e.g., when driving with intelligent speed adaptation, drivers are restricted in their free choice of speed. This may lead to stronger accelerations to the speed limit, since this is the only part that is still under their control. Also, drivers may choose to pass a red light because they feel that time was lost because they cannot drive as fast as preferred, trying to save some time this way.

To exemplify further possible negative system effects, it can be useful to consider the results of a study on behavioral adaptation to Intelligent Speed Adaptation (ISA).

The relationship between speed and accident risk is well known. Therefore, it seems reasonable to introduce safety systems that restrict driving speeds. Studies of ISA generally show considerable reduction in accident risk. However, a study of ISA in Finland indicates that behavioral changes may take place when driving with ISA that negatively related to safety when used under snowy and icy conditions with slippery road, jeopardizing safety (Peltola and Kulmala 2000). This adaptation took place since the type of ISA studied gave feedback to the driver about the current driving speed compared to the officially posted speed, but did not include reduced speed advice due to slippery road. The study, therefore, showed that without dynamic feedback on road conditions (in this case with ISA), drivers drove faster than they otherwise would under such conditions without ISA.

This observation of behavioral adaptations to ISA among Finnish drivers may involve several types of underlying changes. For example, a change in attitude (i.e., overreliance/shift in locus of control) may lead to a change in driver state (inattentive). The changes in attitude may lead to perceptive changes (not seeing, feeling the slippery icy road as dangerous). This is possibly linked to cognitive factors (e.g., not comprehending the limitations of ISA).

Behavioral changes, i.e., driving too fast on icy roads and faster than non-equipped is also demonstrated in earlier driving simulator studies (Comte and Jamson 1998) that found similar changes to driving with ISA in fog. As long as ISA speed limits are fixed and not variable, the combination of ISA and low friction warning is necessary and sufficient to avoid such negative behavioral adaptation, as demonstrated in the INTRO project (Kircher and Thorslund 2009).

4 When Does BA Occur?

Of course, it is very interesting to understand when behavioral adaptation will occur. The first item that needs to be changed in order for BA to take place is a change in the road-vehicle-user system. In case there are no changes in this cooperation, there will not be any behavioral adaptation process. A second precondition for behavioral adaptation is that the feedback can also be perceived. This means that drivers either have to notice the positive or negative effects themselves, or they receive information about the expected positive or negative effects of the change.

5 Behavioral Adaptation in a Longer-Term Perspective

Changes in driver behavior may occur shortly after driving with a system (e.g., if a driver is time pressed, intoxicated, bored, etc.) but the behavioral change may also occur on a longer-term basis, for instance as we age or, over time, grow accustomed to ADAS. By driving in new social settings or rarely occurring traffic scenarios, we may learn new aspects of ADAS use or experience new system limitations. It is important to take into account the fact that an effect may not appear immediately when the driving context is changed, but usually appears after a familiarization period. This is important to realize, since an experiment, aiming to study the effects of a safety system by studying driving behavior as response to the system may reveal positive effects, whereas they disappear after longer-term use.

Draskóczy (1994) outlines chronological phases in behavioral adaptation to driver assistance systems, which incorporates the establishing of stability in performance. She suggests that studies should be done (a) before system activation, (b) immediately (within a month) after system activation, and (c) after 6 months of system use. Only then the real safety effects can be studied and insight into the behavioral adaptation can be found. It has been proposed that traffic safety research on ADAS would benefit theoretically, methodologically, as well as scientifically in terms of more valid predictions of safety effects by extending the scope of interest to include behavioral adaptation in a longer-term perspective (Nardi 1996; Draskóczy 1995; Smiley 2000; Carstens and Nilsson 2001; Saad 2006).

► **Table 6.1** summarizes characteristics of the learning phases and the adaptation to ADAS (Jenssen 2010). It is supported by recent experiences from longer-term studies of ISA (Carstens 2008; Berg et al. 2008) and studies of longer-term use of ESC (Rudin-Brown

Table 6.1
Characteristics of five learning phases in the behavioral adaptation to ADAS

Learning phase	Level of experience	Behavior	Duration	Scenario experience	Typical learning	Typical problems
1. First encounter	Tabula rasa	Exploratory	First day <50 km 1–6 h	Limited	Interface use	HMI related – distraction – distrust
2. Learning	Novice	Unstable	3–4 weeks <1,000 km 10–40 h	Most urban, rural road/ traffic conditions including day/night driving	Controllability	HMI related distraction System limitations
3. Trust	Relatively experienced	Relatively Stable	1–6 months	Most urban, rural road types including day/night driving and many weather conditions	Trust Shift in locus of control	Passive monitoring Overreliance Drowsiness
4. Adjustment	Experienced	Stable	6–12 months	All urban rural road types most summer winter conditions	Functional limitations Malfunction	Resentment
5. Readjustment	Expert	Very stable	>1–2 years	All relevant road traffic conditions	Rarely occurring hazard events System limitations and Malfunction	Mistrust Resentment Loss of manual control skills

et al. 2009). There seem to be five learning phases characterized by level of experience, driver behavior, typical learning, and typical problems. Some of the interviewed drivers had driven with ACC and ESC for up to 6 years and could report both sudden and gradual changes in understanding and control of their ADAS equipped vehicles.

Sudden changes do not occur until a change in the combination of driver state/behavior/workload, ADAS configuration, and road traffic condition occurs.

Examples are: long journey with ACC with a driver glancing at newspaper in passenger seat while a dog suddenly crosses the road or that ESC hinders you from getting up a snowy hill in wintertime. Such hazard events are rare, and might first be experienced after 1–2 years of driving or after a winter season.

The *First encounter* phase is characterized by initial learning of the system interface use. Some drivers read instructions carefully prior to use, but the in-depth interviews and survey results (Jenssen 2010) indicate that most drivers learn as they drive. The duration of the first phase depends on how self-explaining and intuitive the HMI solution is.

The second phase *Learning* is better documented in the literature and typically has a duration of 3–4 weeks. The duration of the learning phase may, however, vary to some extent depending on the type of ADAS studied. How intuitive the ADAS and respective HMI solution is, may play an important role. For example, the ACC system requires system input of set speed and distance to lead vehicle and an understanding of the interface, while ESC requires no input from the driver. ACC may thus take longer time to learn. Yet, optimal effect of, e.g., ESC is first achieved if the driver knows how it works best, i.e., use of full brake force. It is not given that all drivers will learn this by trial and error. Some minimum education or training on ADAS function might be required to optimize safety benefits.

After behavior reaches some kind of stability (phase 3) ADAS use is characterized by system *trust*. A shift in locus of control from driver to vehicle system may develop in this period. However, if there is no trust in the system, this phase 3 will not be entered and the driver will inactivate the system. Related problems are typically overreliance, passiveness, and drowsiness.

Phase 4 involves *adjustment* of trust as some of the most frequent system limitations are revealed. Resentment against system use may evolve and surface in this phase. The learning phases 3–5 seem to be more dependent on scenario experience than kilometers driven. If you mainly drive the ADAS equipped vehicle back and forth on the same route (e.g., to work), you do not necessarily learn relevant functional limitations that are scenario dependent.

Functional limitations and malfunctions may be learnt first after a winter season or after leisure time related to driving in more diverse and novel surroundings for ADAS use (Adjustment phase).

Readjustment of behavior (phase 5) may require as much as 1–2 years of ADAS driving since this readjustment is based on rarely occurring hazard events which are revealed only when a certain combination of road traffic conditions occurs. Mistrust as well as loss of manual control skills may typically surface in this phase.

During the learning process behavior changes from effort-demanding controlled behavior (e.g., learning to use and understand HMI interface and system function) to effortless automatic behavioral control. The novice driver uses conscious problem solving to implement actions while for the experienced driver action control is driven by expectations.

Incidents or accidents in the first two phases can be related to spending too much time and effort on HMI tasks or errors in setting of system parameters, while breach of expectations related to ADAS function can typically occur at the *trust* and *readjustment* phase.

This has implications for methodological challenges in the study of ADAS and interpretation of results from short term versus longer-term exposure to ADAS.

6 Relation Behavioral Adaptation and Safety

As indicated, BA affects safety. In some cases, this adaptation can be positive in terms of safety (e.g., a driver slowing down while making an important business call), or negative. Since this book focuses on intelligent vehicles and driver support systems, the behavioral adaptation in that case will mostly be decreasing the originally aimed for safety benefits.

However, we should keep in mind that BA does not totally eliminate the effects of the safety measures. How much the safety effects are reduced by means of behavioral adaptation is unknown, since it will depend on the type of ADAS, the design of the ADAS, the driver, the current state of the driver, and the local traffic and weather conditions. It is not yet clear whether BA always happens, or what would distinguish cases in which they do from cases in which they do not.

To come to terms with these questions we would need up-to-date valid and quantitative models of road user decision making. Elementary utility models (O'Neill 1977; Janssen and Tenkink 1988) have already paid some services in this respect. In the Janssen and Tenkink model (see Fig. 6.3), the road user is assumed to balance the (dis)utilities of time loss during the trip, plus the possible accident risk, against the utility of being at the destination. From this a choice of optimal speed, and possibly of other driving behavior parameters, then follows so as to be at the optimum of that balance. It has been derived from this type of consideration, e.g., that a device that has an expected effectiveness (i.e., an engineering estimate) ε will not reduce accident risk with that same factor but with a factor that happens to be

$$\hat{\varepsilon} = 1 - (1 - \varepsilon)^{-1/(c+1)}$$

in which c is Nilsson's (Nilsson 1984) parameter in his speed-risk function, which has values of between 3 and 7 for different types of accidents. For fatalities, $c = 7$. It is clear that the safety effect to be realized will always be less than the expected effectiveness: see Fig. 6.4.

For example, with a commonly expected safety effect of $\varepsilon = 0.43$, the estimated effect to be achieved for the fatality rate per kilometer would be in the order of 7% rather than 43% (at 100% use rate).

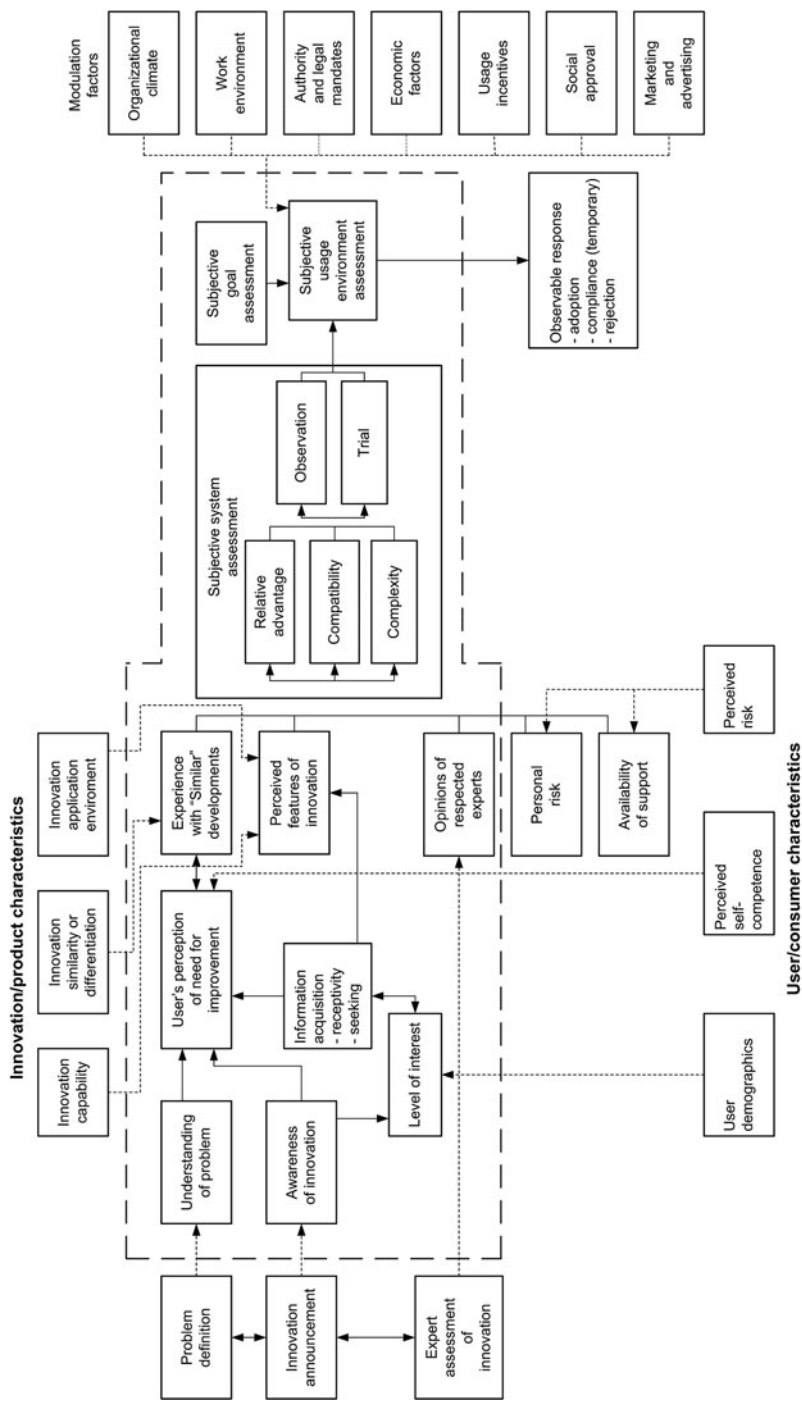
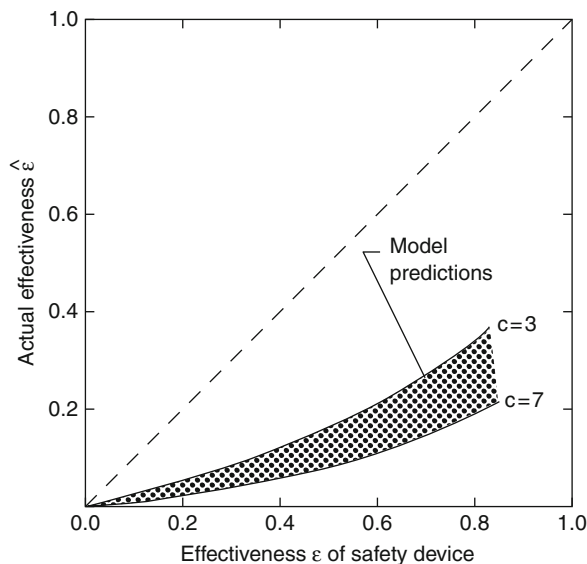


Fig. 6.3

Utility model (Janssen and Tenkink 1988) shows how drivers select optimum speed as a function of time (opportunity) losses and accident risk, so as to make the resulting total expected loss minimal. It appears to be generally true that whenever accident risk is objectively reduced ("after" situation) the optimum speed that is selected will move toward the higher end of the scale



■ Fig. 6.4

Expected and actual (i.e., pre- or postdicted) safety benefit, according to a simple utility model of driver behavior

In order to have proper ex-post safety estimations of safety systems, there need to be a detailed accident analyses (most of the time they are not available) and in-detail comparisons between drivers using these systems and drivers not using these systems. In order to get better safety estimates, Field Operational Tests are being implemented, analyzing the data from many road users with various systems under normal and realistic driving conditions (for more information on FOTs, see the EuroFOT project from the European Commission, <http://www.eurofot-ip.eu/>).

7 Behavioral Adaptation and Acceptance

One important element in the context of behavioral adaptation is “acceptance.” Intelligent in-vehicle systems are introduced to improve throughput, safety, emissions, or driver comfort. Therefore, user acceptance of any system is of major importance. If a system will not be accepted, it will be switched off, drivers will bypass the system, and it may be distractive and counterproductive. However, user acceptance does not always lead to good results, e.g., because drivers may overrely on the system (in case of good acceptance) or show more risky driving behavior due to the subjective feeling of safety. This directly relates to behavioral adaptation.

The issue of user acceptance of in-vehicle technology is a complicated one, since acceptance is not a status quo. People may state on forehand that they will accept a certain system. However, after using the product for a while, they no longer do. The

other way around, driver acceptance may improve after training. Therefore, it is important to understand the aspects that lead to user acceptance when thinking about BA.

In general, we can state that acceptance depends on the following aspects:

- Relative advantage of having the system
- Apparent complexity of the system
- Ease of use
- Compatibility with driving activities
- Safety improvements (subjective)
- Relative importance of the system for driver
- Relative personal risk
- Costs
- Trust in the system
- Reliability of the system

By looking at this list, one can see that a proper design, taking the future user as a starting point, is of major importance when designing an acceptable system. Ergonomics is the basis for the design of any system. However, after a system has been developed, it depends on the type of system, its reliability and experienced use how acceptable a system is and what the behavioral consequences are.

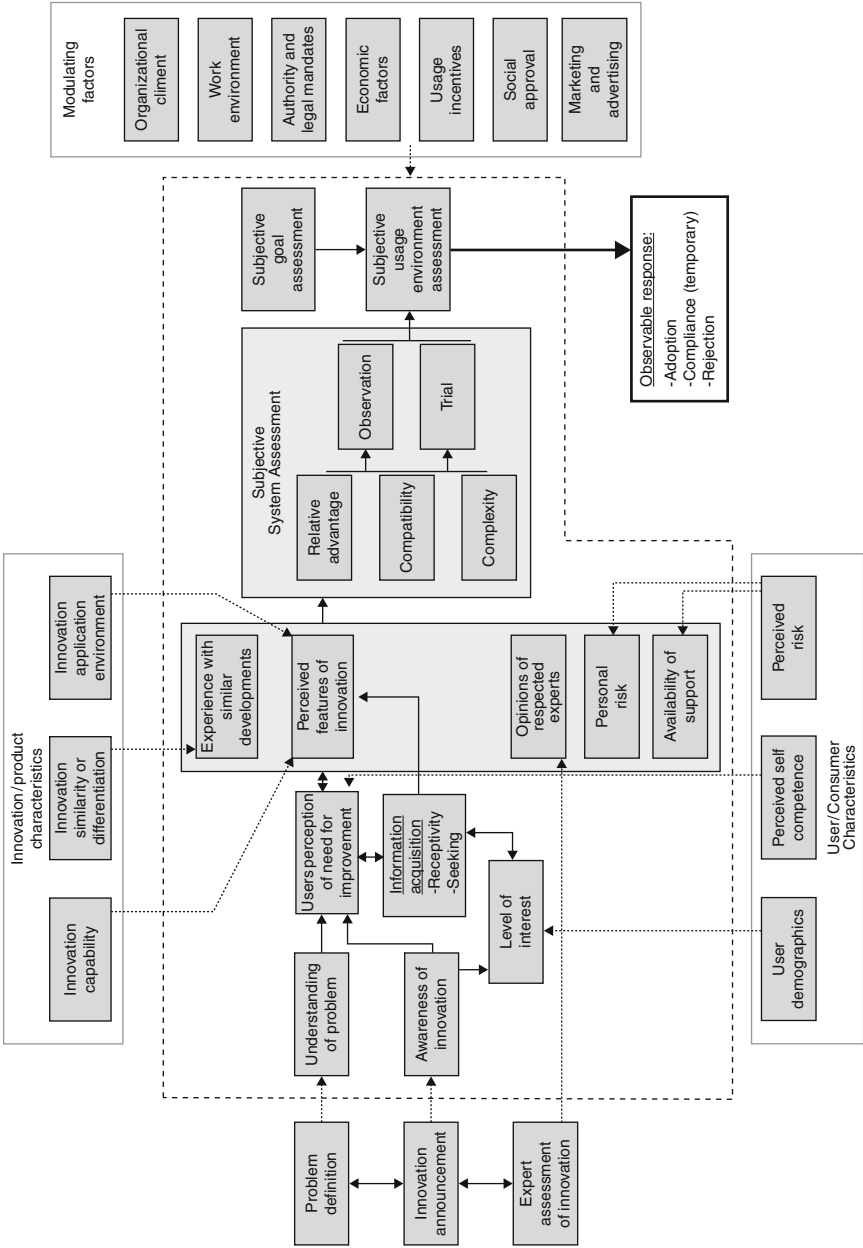
In this, the ratio between hits and misses/false alarms (see also [Table 6.2](#)) is of utmost importance. If a driver is warned for something that was not experienced to be dangerous (false alarm), the system will lose credibility points, decreasing acceptance. The same holds for misses; In case the system does not warn for something that was indeed experienced to be dangerous, the system will not be accepted.

A structural model (see [Fig. 6.5](#)) of the components of innovation acceptance is based on Mackie and Wylie (1988, also described in Kantowitz et al. 1997). The model clearly shows that after understanding the problem and the feeling for the need for improvement, past experience determines the subjective assessment of the system. In this stage, relative advantage of the innovation, compatibility, and complexity of the observation (empirical data) and trials (hands-on experience with the system) determine the assessment. Subjective goal assessment refers to the assessment of factors that are related to the adoption process (e.g., travel time). Subjective usage environment assessment determines the response since it makes the trade-off between the negative and positive results of all factors.

Table 6.2

Explanation of hits, false alarms, misses, and correct rejections. The number of misses and false alarms should be as low as possible and the number of correct rejections but especially the hits should be as large as possible

Warning/situation	Dangerous	Non-dangerous
Warning	Hit	False alarm
No warning	Miss	Correct rejection



■ Fig. 6.5
Mackie and Wylie's model of innovation acceptance (Adapted by Kantowitz et al. 1997)

There are many factors that can influence the acceptance of a driver support system. If the system works fine but the price is extremely high, the innovation may not be accepted in the end. Also, if it is not the driver's free will but rather forced or obliged to be used, acceptance may be low. If there is a privacy risk involved or the personal freedom is limited, acceptance may be low. Also, people have to be aware of the new technology before it can actually be accepted.

8 Future Directions

An important issue regarding ADAS development is whether we should aim to automate the driving task. In principle, when automating the driving task, a large accident source is eliminated, that is the human error due to impairment, mistakes, lapses, or violations. If we automate the driving task, with the driver completely out of the loop, behavioral adaptation is no longer possible, so will this save all our problems?

Speed control is a key component in emotional aspects of driving, related to pleasure, joy of driving, learning to master driving skills, and control of the vehicle. The car is a symbol of modern freedom and mobility with a strong links to self-esteem and life quality. The increasing automation of driver tasks with the introduction of ADAS raises general issues of new automation, all related to acceptance, such as mistrust, resentment, and resistance experienced in other areas of technological development. For drivers with exempted license due to physical, visual, or cognitive impairments, ADAS and robotic vehicles may afford new means of individual mobility.

Robotic crash free electric vehicles were demonstrated by engineers and scientists in 2009 as part of the European project, CITYMOBIL. This was a demonstration of state of the art driverless, crash free, and environmentally friendly vehicles, e.g., shuttling back and forth between a county hospital and surrounding car parks and mass transport terminals. Driverless vehicles are not any longer science fiction, but an end product of a long line of European projects like ADASE, Chauffeur, Netmobile, Cybermove, etc., aiming to develop so-called cybercars, cybertaxis, and other automated cybertransport for personal or public transport. These types of transport function are to a large extent like an elevator, except on the horizontal plane instead of the vertical one. The passenger control is absent or limited to pushing a button to select the desired destination much similar to user control of an elevator. Robotic vehicles so far have proven to operate safely at low speeds (max 30–40 km/h), but this may in the future be extended to speed levels suitable for motorway driving on dedicated lanes. According to researchers, there are still many barriers to high speed driverless vehicles and especially for safe high speed transport on the open road network (Wahl et al. 2007).

According to the development roadmap within the PREVENT project (Sjögren 2008), quite some effort has so far been on longitudinal and lateral control of individual vehicles. Integrated safety between cars and infrastructure (i.e., Car to X safety) is aimed at the year 2018, with the ultimate goal of developing Cooperative integrated safety systems. Parent and Yang (2004) envisage a divided road traffic network; one network for automated driving and one network for manual control of vehicles. Dual mode vehicles may be

developed in order to operate on both types of network. Hence, an important research challenge will be to study how drivers cope with the transition from one mode of driving to the other. The scientific challenge lies not only in identifying barriers to deployment, and overcoming resistance to automation, but also in research issues related to loss of manual control skills.

In the current transport scenario where vehicles with and without ADAS have access to the road traffic system, habitual ADAS drivers will experience an increasing gap between skills required for navigation, maneuvering, and control of new vehicles with ADAS, compared to normal cars still relying on manual control of all driver tasks. Drivers may lose the cognitive maps of a route we today learn and develop through mental notes of landmarks, nodes, etc. Will ADAS selective drivers in the future experience problems when trying to find their way in, e.g., London when occasionally using a vehicle without navigation aid? Will intervention be less effective or even hazardous in the future when driving in an unassisted mode or if ADAS, automatically controlling lateral and longitudinal maneuvering suddenly malfunction? Navigation support may relieve the driver and give more resources to the traffic situation, but malfunction and misguidance have already led to accidents as reported in several newspapers.

ADAS can already park the car for you (Automatic Parking), take you safely down a steep hill (Hill Descent Control), automatically brake the car if a minimum safe distance to the car ahead is overridden (ACC), control speed (ISA), automatically stabilize the vehicle in a skid (ESC), and warn you every time a vehicle is in your blind spot (LCA). Will observation skills and vehicle control skills deteriorate for habitual ADAS drivers – and at what rate?

Young drivers, 16–24 years old, have the highest accident risk among all age groups. These drivers may also have the largest potential benefit from safety-related ADAS technologies with low opportunity for negative behavioral adaptation since there is no reference driving behavior yet. The largest safety benefit, related to behavioral adaptation is expected for systems for which the safety benefit is not easily perceived by the driver. Conscious speed violations and maladapted speed in road traffic situations due to risk proneness or inexperience is a frequent accident cause among young drivers, challenging further research and development related to speed controlling ADAS for this group of drivers. Designing ADAS with enhanced user acceptance for this age group is therefore important, unless there is political consensus to implement mandatory controlling solutions, such as completely taking the driver out of the loop.

New generations of drivers learning to drive vehicles with ADAS will have increasingly less experience with manually controlled vehicles. How to cope with this fact is a challenge for transport authorities and the research community. This may have a future impact on driver education, licensing, liability, and traffic regulation. Choice of action in this respect should be based on scientific knowledge.

Other points of interest are the fact that automated driving will only be for specifically dedicated lanes, and the fact that sensors are not as intelligent as human drivers in anticipating complex interactions with other road users. Where the human driver recognizes a child throwing a ball and anticipates that it will run into the street to get it, this is

a complicated problem for an intelligent vehicle. Also, the transitions from automated driving to manual driving may lead to unsafety as well as most probably new type of errors.

9 Will New Skills Appear?

Future ADAS research should not only focus on loss of skills but also study how drivers acquire new skills and study new models of vehicle functionality, operation, and use with ADAS.

Anyone who has been introduced to a new car with different size and shape than their former car has experienced the problems of parking in a tight spot when you are unfamiliar with vehicle length. As we gain experience about the outreach of a vehicle and the necessary safety margins needed, this knowledge becomes part of our internal model of the vehicle and procedures for car parking. Hence, car parking can be performed more swiftly and accurately with learning. Even though the actual tip of the front or tip of the rear end is hidden from sight, we “know” where it is. Recent developments within cognitive science and neural correlates of skills confirm that we develop not only abstract cognitive models for objects but also develop connections on a neurological brain level representing this knowledge.

Studies with functional neuro-imaging have demonstrated the brain areas representing the tip of a stick, but not yet the corner of a car or the skill of handling it (Povinelli 2000; Johnson-Frey 2004). Functional neuro-imaging may in the future represent a convergent validation between behavioral investigation and imaging techniques, showing that the brain correlates conceptual knowledge and complex real world behaviors like driving cars equipped with ADAS. Neuro-imaging may also demonstrate the loss of concepts and skills.

Developments within automotive safety systems have already given and will continue in the future to contribute to a considerable reduction in road traffic fatalities. It is one of the most important potential contributors to increased traffic safety, especially in countries where traditional safety measures like road design, education, and surveillance are already exploited. With the current renewal rate of more international cars, car models from before 2000 will be renewed first in 2017. Sakshaug and Moe (2006) argue that in a 15 year perspective, incentives leading to a reduction in car renewal by 5 years may imply a reduction of 250 traffic fatalities just due to the improved passive safety in modern vehicles we had up to 2004 models. Potential reduction in traffic fatalities due to ADAS comes in addition.

10 International Aspects of Adaptation to ADAS

International differences related to user effects and the potential safety impacts of ADAS are important to establish. One good example related to international aspects of ADAS is eCall. eCall is a project of the European Commission intended to bring rapid assistance to motorists involved in a collision anywhere in the European Union. The eCall system

developed employs a hardware black box installed in vehicles that wirelessly sends airbag deployment and impact sensor information, as well as GPS coordinates to local emergency agencies.

Virtanen et al. (2006) studied in-depth fatal accident data in Finland, covering 1,180 fatalities of which 919 were motor-vehicle occupants. Time delay between accident occurrence and notification of the emergency response center was calculated. Two trauma specialists evaluated whether a fatality would have been prevented had there been no delay in accident notification. The results showed that eCall could have prevented approximately 4–8% of the road fatalities that occurred in Finland during 2001–2003.

This result may not be transferable to all other countries. eCall is likely to be more effective on remote roads with low levels of traffic. A major part of the road network in Finland fits this description as well as northern parts of Norway. A seriously injured driver may in worst case lie unheeded for hours or days if involved in an accident in remote parts of the country. In comparison, eCall in densely populated countries like the Netherlands may have little or no effect. Accidents are immediately spotted and reported on a majority of the road network by other road users. Hence traffic density and the makeup of road network may have an important impact on the safety effect of a system like eCall.

The quality of emergency services may also play an important role. eCall will evidently be more effective in countries where emergency services are well organized, and delays in response following emergency call receipt are minimal.

International aspects of adaptation to ADAS can also be related to increased mobility. For example, there has been an increase in foreign freight transport and foreign professional drivers on Scandinavian roads with the open market policy in EU drivers who are unfamiliar with the Nordic road traffic conditions. With increasing use of ADAS in heavy vehicles there will be an increasing number of professional drivers who may encounter winter related functional limitations and malfunction of ADAS for the first time. The consequences of ADAS-related incidents and accidents may be serious with vehicles up to 50 t traveling on icy narrow roads not designed for this type of traffic, i.e., heavy vehicles with ESC not able to get up hills.

11 General Conclusions

That behavioral adaptation will occur with the introduction of ADAS is predictable – we should be more surprised if there is no sign of it. The nature and direction of these adaptations, intended or unintended, positive or negative, are important to verify in order to assess possible impacts on road safety.

How to prevent BA from occurring in relation to the introduction of ADAS is yet not fully understood. Based on our knowledge of conditions where BA is likely to occur, suggestions for further research and development of ADAS can be done. Some examples are research into the design and use of unperceivable or non-obtrusive measures, the use of measures that give little freedom of action, and the use of additional warnings to avoid errors based on malfunction, overreliance, or misunderstanding of system function.

Ideally, such warnings should be unnecessary if all aspects are considered and built into system functionality from early in the design process, yet it is human to err (even on the designers side) and it is likely that we have to search for ad hoc solutions to problems of behavioral adaptation also in the future. As long as designers take the presence of BA, and its potential to reduce the engineering estimate of safety, into account, a lot is already won. The next decades of research have to show the reducing effects of different types of ADAS.

References

- Berg C, Bayer SB, Thesen G (2008) Ung trafikk. Resultater fra ISA forsøk med unge førere i Karmøy, in Norwegian (results from ISA study with young drivers in Karmøy). Rapport-IRIS 2008/149, Stavanger
- Bjørnskau T (1994) Spillteori, trafikk og ulykker: en teori om interaksjon i trafikken. In: Norwegian (Game theory, traffic and accidents: a theory on interaction in traffic) Doctoral thesis, University of Oslo, Norway
- Cacciabue PC (ed) (2007) Modelling driver behaviour in automotive environments. Springer, London
- Carsten OMJ (2001) From behavioural adaptation to safety modelling: predicting the safety impacts of new technologies. Behavioural Research in Road Safety XI. Department of transport, local government and the regions, London
- Carsten O (2008) Estimating the effects of ADAS introduction on Safety: effects of scenario and system. In: Safe highways of the future. Conference proceedings, Brussels
- Carsten OMJ, Nilsson L (2001) Safety assessment of driver assistance systems. Eur J Transp Infrastructure Res 1(3):225–243
- Comte S, Jamson H (1998) The effects of ATT and non-ATT systems and treatments on speed adaptation behaviour, deliverable D10 in the MASTER project. In: VTT, Espoo
- Draskóczy M (ed) (1994) Guidelines on safety evaluation. In: DRIVE Project V2002 – Lund University, Lund
- Draskóczy M (1995) Guidelines on safety evaluation of transport telematics systems. DRIVE Project V2002, HOPES
- Elvik R, Vaa T (2004) The handbook of road safety measures. Elsevier Science, Oxford
- Engström J, Hollnagel E (2005) A general conceptual framework for modelling behavioural effects of driver support functions. In: Cacciabue PC (ed) Modelling driver behaviour in automotive environments. Springer, London, pp 149–164
- Fosser S, Sagberg F, Sætermo IA (1997) An investigation of behavioural adaptation to airbags and antilock brakes among taxi drivers. Accid Anal Prev 29(3):293–302
- Hollnagel E, Woods DD (2005) Joint cognitive systems: foundations for cognitive systems engineering. CRC Press, Boca Raton
- Janssen WH, Tenkink E (1988) Considerations on speed selection and risk homeostasis in driving. Accid Anal Prev 20:137–143
- Jenssen GD (2010) Behavioural adaptation to advanced driver assistance systems. Steps to explore safety implications. Doctoral Thesis at NTNU 124. ISBN 978-82-471-2217-4
- Johnson-Frey SH (2004) The neural bases of complex tool use in humans. Trends Cogn Sci 8(2):71–78
- Kantowitz BH, Lee JD, Becker CA, Bittner AC, Kantowitz SC, Hanowski RJ, Kinghorn RA, McCauley ME, Sharkey TJ, McCallum MC, Barlow ST (1997) Development of human factors guidelines for advanced traveler information systems and commercial vehicle operations; exploring driver acceptance of in-vehicle information systems. <http://www.fhwa.dot.gov/tfhrc/safety/pubs/96143/index.html>
- Kircher K, Thorslund B (2009) Effects of road surface appearance and low friction warning systems on driver behaviour and confidence in the warning system. Ergonomics 52:165–176
- Mackie RR, Wylie CD (1988) Factors influencing acceptance of computer-based innovation. In: Helander M (ed) Handbook of human-computer interaction. Elsevier, New York, pp 1081–1106
- Michon JA (1985) A critical view of driver behaviour models. What do we know, what

- should we do? In: Evans L, Schwing R (eds) *Human behaviour and traffic safety*. Plenum, New York, pp 485–525
- Mourant RR, Rockwell TH (1970) Mapping eye movement patterns to the visual scene in driving: an exploratory study. *Hum Factors* 12:81–87
- Nardi BA (ed) (1996) *Context and consciousness: activity theory and human-compute interaction*. MIT, Cambridge
- Nilsson G (1984) Speeds, accident rates and personal injury consequences for different road types. Linköping, VTI Rept. 277, Sweden
- O'Neill B (1977) A decision-theory model of danger compensation. *Accid Anal Prev* 9:157–165
- OECD (1990) *Behavioural adaptations to changes in the road transport system*. Organization for economic co-operation and development, Paris
- Parent M, Yang M (2004) Road map towards full driving automation. In: *Proceedings of ITS world conference*, Nagoya CD ROM
- Peltola H, Kulmala R (2000) Weather related ISA – experience from a simulator. In: *Proceedings of the 7th world congress on intelligent transport systems*, Turin, 6–9 Nov 2000
- Peltzman S (1975) The effects of automobile safety regulations. *J Polit Econ* 83:677–725
- Povinelli DJ (2000) *Folk physics for Apes: The Chimpanzee's theory of how the world works*. Oxford University Press, Oxford
- Rudin-Brown CM, Noy YI (2002) Investigation of behavioral adaptation to lane departure warning. *Transp Res Rec* 1803:30–37
- Rudin-Brown CM et al (2009) Does electronic stability control change the way we drive? In: *Proceedings of the 88th Annual Meeting of the Transportation Research Board TRB*, Washington DC, 11–15 Jan 2009
- Saad F (2006) Some critical issues when studying behavioural adaptations to new driver support systems. *Cogn Tech Work* 8:175–181
- Sakshaug K, Moe D (2006) TS-tiltak frem mot 2020: Nye biler redder liv! In: *Norwegian (Traffic safety measures towards 2020. New cars save lives)*. Samferdsel nr 1 Institute of Transport Economics, Oslo
- Sjøgren A (2008) Creating a cost effective integrated safety system – INSAFES towards integration. *Safe highways of the future*, Bruxelles www.prevent-ip.org. Accessed 12 Feb 2008
- Smiley A (2000) Behavioral adaptation, safety, and intelligent transportation systems transportation research record 1724, issue 00-504, pp 47–51
- Vaa T, Gelau C, Penttinen M, Spyropoulou I (2006) ITS and effects on road traffic accidents. State of the art. In: *Proceedings of the 13th world congress on intelligent transport systems*. London
- Virtanen N, Schurokoff A, Luoma J, Kulmala R (2006) Impacts of automatic emergency call system on accident consequences. Ministry of transport and Communications, Helsinki, Finland
- Wahl R, Tørset T, Vaa T (2007) Large scale introduction of automated transport. Which legal and administrative barriers are present? ITS for a better life. In: *Proceedings from the 14th ITS World Congress*, Beijing
- Weller G, Schlag B (2004) Verhaltensadaptation nach Einführung von Fahrerassistenzsystemen. In: Schlag B (ed) *Verkehrspsychologie mobilität – Verkehrssicherheit – Fahrerassistenz*. Pabst Science, Lengerich, pp 351–370
- Wilde GJS (1982) The theory of risk homeostasis: implications for safety and health. *Risk Anal* 2:209–225
- Wilde GJS (1988) Risk homeostasis theory and traffic accidents: propositions, deductions and discussion of dissension in recent reactions. *Ergonomics* 31:441–468
- Wilde GJS (1994) Risk homeostasis theory and its promise for improved safety. In: Trimpop RM, Wilde GJS (eds) *Challenges to accident prevention: the issue of risk compensation behaviour*. Styx, Groningen

7 Simulation Approaches to Intelligent Vehicles

Bart van Arem¹ · Martijn van Noort² · Bart Netten²

¹Civil Engineering and Geoscience Transport & Planning, Delft University of Technology, Delft, CN, The Netherlands

²Netherlands Organization for Applied Scientific Research TNO, Delft, The Netherlands

1	<i>Introduction</i>	141
2	<i>Simulation Environment</i>	142
2.1	The Role of Simulation and Modeling in the Development Process	142
2.2	Requirements for Simulation and Modeling	144
2.3	Simulation Environment	145
3	<i>Traffic Flow Simulation Models</i>	147
3.1	The Basics of Traffic Flow Simulation	147
4	<i>Concept Development</i>	150
4.1	Introduction	150
4.2	Cooperative Adaptive Cruise Control	150
4.3	Simulation Setup	151
4.4	Results	151
4.5	Discussion	152
5	<i>Application Testing and Verification</i>	153
5.1	Introduction	153
5.2	Requirements on Simulation in a Laboratory Test Environment	153
5.3	An Example: The SAFESPOT Test Bench	154
5.3.1	Setup 1: Application Unit Testing	155
5.3.2	Setup 2: Testing the Integration of Applications with Facilities	155
5.3.3	Setup 3: Application Platform Testing	157
5.3.4	Setup 4: Testing the Cooperation with External Vehicles and Roadside Units	157

6 *Application Validation and Evaluation* 158

6.1 An Example: Shockwave Damping 158

7 *Conclusions and Challenges* 161

Abstract: The development of systems for intelligent functions is a very complex process, involving many technological components, requirements with respect to robustness and fail safety, time pressure, and different stakeholders. Simulation plays an important role in the development process and can support the development process from the early concept development, through testing and verification of components to the validation and evaluation of systems. Requirements and examples of simulation environments needed are discussed. Traffic flow simulation is used to show how it can be used in the development of the concept of Cooperative Adaptive Cruise Control. Next, simulation is applied for testing and verification of cooperative safety applications. Finally, simulation is applied for validation and evaluation of applications tested in a Field Operational Test.

1 Introduction

There is a worldwide increasing interest in the application of intelligent systems in road vehicles. It is believed that intelligent systems in vehicles that support or even take over driving tasks can enhance driving comfort and trip quality. Intelligent systems such as Adaptive Cruise Control, Lane Keeping Assistance, Collision Avoidance and Dynamic Navigation can lead to more reliable and efficient traffic flows, more comfort, improved traffic safety, and reduced fuel consumption.

The increasing interest in intelligent vehicles is also driven by the tremendous advances in sensors, sensor networks, communication technology, computing power, and controllers. The application of intelligent functions in vehicles is also driven by the need for product innovation, using information and technology to produce comfortable, safer, and cleaner vehicles. In particular, the use of communication technology opens the door to a generation of new products and services based on communication between a vehicle with other vehicles and roadside systems.

Intelligent systems in vehicles and the use of communication with sensors as well as sources outside the vehicles lead to increasing complexity in the development process of intelligent vehicles. In addition, there is a need for robust and failsafe systems as intelligent systems can partly or entirely take over the execution of driving tasks from the driver. At the same time, there is an increasing pressure to develop effective and affordable systems in a shortening duration of the development process. Finally, as intelligent systems use components and information from a variety of suppliers, many actors are involved in the development of systems for intelligent systems in vehicles, ranging from car manufacturers, suppliers, road operators, to traffic industry.

The development of intelligent systems is a very complex process, involving many technological components, requirements with respect to robustness and fail safety, time pressure, and different stakeholders. Simulation plays an important role in the development process and can support the development process from the early concept development, through testing and verification of components, to the validation and evaluation of systems. Simulation can be used to test components and systems fast and safely in a virtual

environment and in a wide range of hypothetical scenarios. Simulation can significantly reduce the uncertainty about the technical performance and the expected effectiveness of the systems in field tests.

Simulation also plays a vital role in the evaluation and validation of the developments of intelligent functions. Policy makers and road authorities have a need for the evaluation of intelligent systems in vehicles. An evaluation of the societal impacts of these functions helps in the formulation and assessment of policies for improving mobility. Road authorities are faced with the decision whether to invest in traditional roadside systems for traffic management and information or in more innovative but less well-known in-vehicle systems. This decision needs an assessment both of the technical and the societal aspects of new technologies. Simulation plays an important role in these decision processes, being the prime tool for investigating future scenarios and “what-if” questions.

This chapter focuses on the simulation approaches to intelligent functions in vehicles. Throughout this chapter, the following terms are used:

- A *function* defines the purpose of a system that can be defined in a model, mathematical relation, or controller such as a Cooperative Adaptive Cruise Controller (CACC). The term “function” is typically used in concept development, where the behavior and structure of a system are not yet considered.
- An *application* defines the realization of a function in application software. A basic set of applications are defined in (ETSI 2009) that run as software components in an application layer for ITS systems (ETSI 2010).
- A *system* defines a collection of components organized to accomplish a specific function or set of functions. A system takes a holistic view, including not only the components in hardware and software, but also the users, networking, communication, processes, function, behavior, and structure.

This introductory section has explained the importance of simulation in the development process of intelligent vehicles. ➤ Section 2 will focus on the simulation environment needed. ➤ Section 3 focuses on traffic flow simulation as it is key in the overall simulation environment. ➤ Section 4 shows how simulation can be used in the development of the concept of Cooperative Adaptive Cruise Control. In ➤ Sect. 5, simulation is applied for testing and verification of cooperative safety applications. In ➤ Sect. 6, simulation is applied for validation and evaluation of applications tested in a field experiment. Finally, the chapter closes with conclusions and an outlook in ➤ Sect. 7.

2 Simulation Environment

2.1 The Role of Simulation and Modeling in the Development Process

Simulations and models are developed and applied in various ways throughout the development of intelligent functions. This section identifies three stages in the

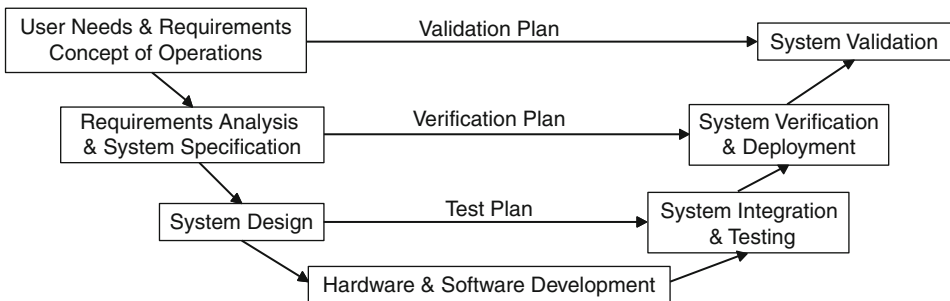
development process where the role of simulation is most obvious: concept development, application testing and verification, and application validation and evaluation.

Development of intelligent and novel functions is a risk-full, long-term, and costly process. Many concepts have been developed to some level of maturity and never made it into a successful product. Methodologies are developed to manage the risks and investments. Technology Readiness Assessment will be used as an example here. This methodology is based on the concept that the maturity of a technology can be assessed and scored in Technology Readiness Levels (TRL) (US-DOD 2009). The successive levels can be broadly grouped into the following successive phases of concept, technology, and system and product development.

- Concept Development is the basic research process to develop a concept of the intelligent function with artifacts like theoretical principles, mathematical models, interaction protocols for cooperative systems, control algorithms, and the concept of operations for the intelligent function.
- Technology Development is the applied research process to develop the innovative technology that is key to enabling the realization of the concept, such as new sensors, communication protocols and devices with the required performance.
- System Development and Demonstration is the process to develop a complete system with hardware and software components, and integrate these in a test-vehicle. The functionality and maturity of the system can be demonstrated, verified, and validated on test tracks or on the road.
- Product Development is the final development phase, starting with the development of prototypes of products, and finally with products, maintenance, and product updates.

Each development phase, or more specifically each TRL, aims at the development of the most critical technologies that must be achieved before development at the next level becomes feasible. This results in an iterative development process where each iteration implements some form of the system engineering V-model (ISO 2004) as depicted in Fig. 7.1.

The vertical levels identify the levels of detail in an iteration. The first step defines the user needs and requirements that include the critical technologies. The validation plan



■ Fig. 7.1
System engineering V-model

defines how these needs and requirements have to be tested to validate the feasibility of the critical technologies or the final system. At the second level, the user needs and requirements are analyzed from which the system is specified. The verification plan is defined to verify that the final system is built according to these specifications. At the lower levels, the system and its underlying components are designed according to the specifications, implemented, integrated, and tested.

System engineering takes a holistic view of the system around an intelligent function, including the users, legacy in-vehicle systems, communication systems, roadside and back-office systems to interface with, as well as the processes, behavior, and structure of the system itself. The application here is the realization of the intelligent function at the (OSI 1994) application layer and, more specifically, as an (ETSI 2009) application running on the application or facility layer (ETSI 2010). In this sense, an application is a software component on the application platform of the onboard unit of an intelligent vehicle or other station, using the facilities for communication, local dynamic map, and interfaces to the vehicle network and sensors. This chapter will focus on the use of simulation and modeling for developing and testing the intelligent functions and applications. Supporting technologies and system components will be considered as the environment of such applications.

The role of simulation and modeling is closely related to the maturity of the developed function and application. During the earlier phases of concept and technology development, intelligent functions will be developed and tested in a laboratory environment that includes a simulation environment. Initially the laboratory environment will be fully virtual, meaning that all functions and their environment are simulated. ➤ Section 4 will give a more detailed description and examples of the role of simulation and modeling in concept development.

Gradually, the simulated functions will be replaced by newly developed application components during technology and system development. The laboratory environment has to be adapted to provide the interfaces to the new components and to simulate the behavior of the neighboring functionality at the interfaces of the application or system components. Gradually, the laboratory environment is adapted from a fully virtual environment into a Software-in-the-Loop (SiL) and Hardware-in-the-Loop (HiL) environment. Simulation has a primary role in the laboratory test environment for testing and verification of the applications, as described in more detail in ➤ Sect. 5.

Finally, when the systems are completely integrated in a vehicle, further development and testing of the systems continues in the field. Simulation has a distinguished role in the validation and evaluation of these systems, and later prototypes and products, for example, to scale up the penetration rate of equipped vehicles for impact assessment, as described in ➤ Sect. 6.

2.2 Requirements for Simulation and Modeling

The requirements for simulation and modeling change over the development phases and processes. In particular, requirements on the fidelity of models and simulators will increase with the maturity of the developed function over the successive development phases.

In the concept development phase, the intelligent function itself is subject of research, rather than the behavior or structure of the application or system realizing the function. In this phase, generalized models of traffic will be used, while the effects of sensor and communication systems, human-machine interaction, or weather conditions, for example, may be strongly idealized. A fully virtual simulation environment is typically used in this phase for development and testing. When the concept matures, the model of the intelligent function itself will be refined, and the simulation environment should incorporate models to improve the fidelity of traffic situations. Microscopic traffic models will be introduced with a variety of models for road user behavior, including lateral behavior for overtaking and merging or crossing intersections, and models for human-machine interaction. **Section 3** gives an example of traffic behavior modeling. Other aspects, like sensing and communication will be addressed in more detail in other chapters in this handbook.

In the technology and system development phases, the fidelity of development and test environments increases further from fully virtual, to mixed reality setups with software, hardware, and drivers in the loop. The fidelity of models and simulation environments should increase accordingly. When components are introduced into the vehicle systems, high-fidelity models and simulations for sensors, communication, and vehicle behavior should be introduced. These models should represent the physics and implementation characteristics, as well as noise, accuracy, and reliability flaws. In the final stages of simulation, when performance and operational hazards are developed and tested, implementation details should be introduced in the models like sampling frequencies, processing and communication delays, and asynchronicity of modules.

2.3 Simulation Environment

This section describes the setup of a simulation environment for the three development phases introduced in **Sect. 2.1**. Each phase has its own needs, as discussed in **Sect. 2.2**, leading to different setups for the simulation environment. These will be detailed in **Sects. 4**, **5**, and **6**.

Several kinds of simulation are being distinguished:

- Traffic simulation: simulation of all traffic and traffic-related systems (such as traffic control and traffic management systems) on a network. The purpose is to study the interactions between the actors on the network (vehicles, drivers, traffic management systems, etc.) and the resulting impacts on traffic efficiency, traffic safety, and the environment. Traffic simulation can be microscopic, mesoscopic, or macroscopic, distinguished by the level of detail in which traffic is modeled and the geographic and temporal scale of the simulation.
- Vehicle and application simulation: more detailed than traffic simulation, this type of simulation involves a handful of vehicles and roadside units on a small traffic network component, which are modeled with high-fidelity. HiL and SiL simulation can be included. The purpose is application development or technical evaluation.

- Communication simulation: with the advent of cooperative systems, the need has arisen to simulate telecommunications. The purpose is application development or assessment of communication performance of applications. Communication simulation can be done at various levels of detail depending on needs and available computing power.
- Driving simulators: Many intelligent functions are not autonomous but work in cooperation with the driver. The purpose of driver simulators is to assess driver behavior and the effect on that of the intelligent function.

Each simulation environment has a need for models that describe the dynamics of the actors in the simulator, at a level of detail and accuracy for the simulator in question. For simulating intelligent functions in vehicles, models are needed for vehicle dynamics, intelligent functions (sensors and actuators), drivers, communication performance, and roadside units. Optionally, models can be used for post-processing outputs, such as emission and dispersion models or safety assessment models. Depending on the need, models can sometimes be replaced by dedicated simulators or components. For example, telecommunication in a large-scale simulation is typically handled by a statistical model, but on a smaller scale this model can be replaced by a dedicated communication simulator. Likewise, models for applications can be replaced by HiL or SiL.

Recent trends show a development toward integrated environments that combine several of these tools and models in a single simulation platform, for example, the MOBYSIM platform under development at TNO (van Noort et al. 2010). Depending on the need, different combinations are employed. In the Application Testing and Verification phase (see ► Sect. 5) one can combine vehicle, application, communication, and driving simulators to create an environment where all (local) aspects of the system can be tested and verified. In the Concept Development phase (see ► Sect. 4), a traffic simulator is combined with abstract models. In the Application Validation and Evaluation phase (see ► Sect. 6), the same traffic simulator is used, but with models based on the implemented application. Throughout there is a need to harmonize the models and the simulation world between the various simulators.

Simulation is often used in conjunction with a wider array of R&D tools and tests, including

- Test labs: controlled environments for testing hardware and software in a real (non-simulated) setting. The environment is usually limited in scope. Typical examples are test benches for systems and system components, or larger labs like VeHiL (Verhoeff et al. 2000; Gietelink et al. 2006).
- Closed field labs: semi-controlled environments such as test tracks, where hardware and software can be tested under real environment conditions and with real drivers, but in artificial traffic conditions. The number of scenarios is limited.
- Open field labs: uncontrolled environments such as field operational tests or natural driving studies, where the systems under test are being used by “real” drivers in “real” traffic, with no a priori limitation on the scenarios that can be encountered.

By comparison, simulation can be seen as a tool that allows a high level of control and a large scale in terms of geography, number of actors, and time. It allows the analysis of scenarios that would be unsafe or unfeasible in practice and is relatively low cost. The main challenge is to model the actors in the simulation with sufficient accuracy. There are several solutions to this, such as HiL, SiL, and combining simulation with other tools. These approaches are discussed in more detail in [Sects. 5](#) and [6](#).

3 Traffic Flow Simulation Models

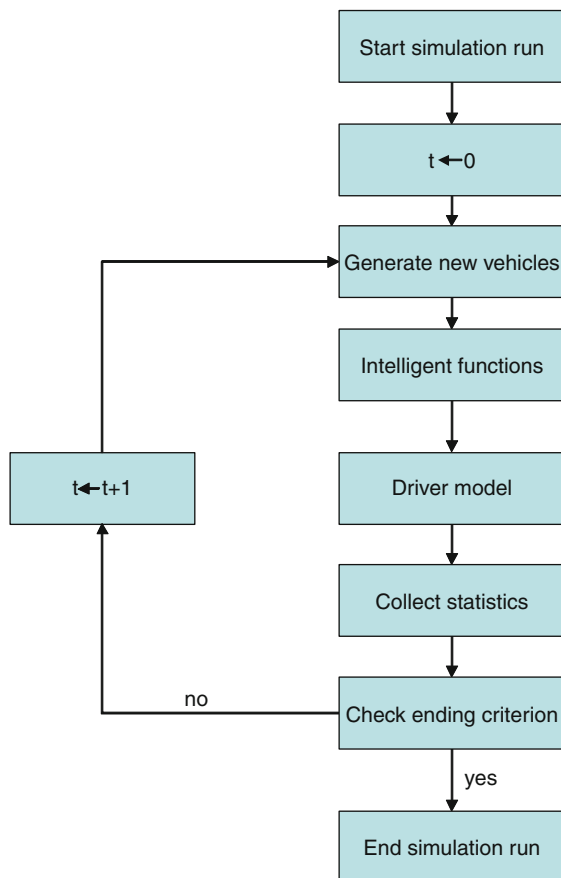
3.1 The Basics of Traffic Flow Simulation

Traffic simulation models are used to study the interactions between vehicles, drivers, traffic management systems on a road network and the resulting impacts on traffic efficiency, traffic safety, and the environment.¹ Traffic simulation models can be microscopic, mesoscopic, or macroscopic, distinguished by the level of detail in which traffic is modeled and the geographic and temporal scale of the simulation. This section describes the basic concepts of microscopic traffic flow simulation models because of their ability to assess the interaction of intelligent functions in vehicles with other vehicles and traffic flow characteristics. The introduction is restricted to the simulation of vehicles as part of a traffic flow on a motorway, without on-ramps or off-ramps. Other aspects such as driver modeling, human-machine interaction, sensors, communication, vehicle dynamics, etc., are described in other chapters of this handbook.

Basically, traffic flow simulation models simulate the movements of vehicle at incremental time steps on a road network. In its basic form, the state space of a traffic simulation models consists of the position (including the lane), speed, and acceleration of each vehicle on the road network. In more enhanced forms of traffic flow simulation, the state space can also include the state of roadside systems, communication devices, etc. The process of conducting a simulation is depicted in [Fig. 7.2](#).

The simulation process is started by initializing the time t at zero. In each time step, the following steps are taken. A random process is used to generate new vehicles at origin nodes of the road network. A model for intelligent functions is applied to observe and interpret the ambient conditions and generate output in terms of information or an automatic action, typically using state estimation and control models. A driver model is used to compute an updated acceleration and lane selection, taking into account the output of the intelligent functions and based on sub-models for route choice, unconstrained driving, car-following, and lane changing. This process is repeated until a stopping criterion, such as a maximum simulation time, is satisfied. This result is used for further data analysis such as speed-volume-density relations, shockwaves, time delays, emissions, etc. By using data from multiple traffic simulation runs, statistical analyses can

¹ © 2011 IEEE. Sections 3 and 4 contain portions reprinted with permission, from Schakel et al. 2010.



■ Fig. 7.2

The process of conducting a simulation run

be done, for example, to generate a confidence interval for the average travel time on the network simulated.

The core of a traffic simulation model is the mathematical specification of the intelligent function and the driver model. The driver model usually consists of a longitudinal and lateral model. The longitudinal driver model results in the longitudinal acceleration of a vehicle. For an extensive review of longitudinal driver models see Brackstone and McDonald (1999). As an example, a modification is described of the Intelligent Driver Model (IDM) which is presented in Treiber et al. (2000). The main feature of the model is the nonlinear response to speed differences, included in s^* , the dynamic desired headway. The acceleration is determined in (7.1).

$$\frac{dv}{dt} = a \cdot \left[1 - \left(\frac{v}{v_0} \right)^4 - \left(\frac{s^*(v, \Delta v)}{s} \right)^2 \right] \quad (7.1)$$

with,

$$s^*(v, \Delta v) = s_0 + vT + \frac{v\Delta v}{2\sqrt{ab}} \quad (7.2)$$

where a is the comfortable acceleration, v is the current speed, v_0 is the desired speed, s_0 is the minimum headway (at standstill), T is the desired time headway, Δv is the speed difference with the leader, s is the current distance headway and b is the comfortable deceleration. The IDM shows realistic shockwave patterns but has a macroscopic capacity of just below 1,900 veh/h, which is too low. In order to reach a reasonable capacity, the desired time headway needs to be lowered to unreasonable values. A modified version of IDM, referred to as IDM+, is based on a minimization over the free-flow and the interaction terms of (7.1), resulting in reasonable capacity values.

$$\frac{dv}{dt} = a \cdot \min \left[1 - \left(\frac{v}{v_0} \right)^4, 1 - \left(\frac{s^*(v, \Delta v)}{s} \right)^2 \right] \quad (7.3)$$

The lateral driving model describes the lane change model. A (voluntary) lane change to the left may be carried out if the following conditions are satisfied:

$$(v(t) \leq k_{vlc}v_0) \text{ AND } (d(t) \leq k_{dlc}s^*) \text{ AND } (d_t(t) \geq d_{lc}) \quad (7.4)$$

With k_{vlc} a factor smaller than one, the first condition expresses that the vehicle is driving slower than the intended speed of the driver. The second term expresses that the vehicle is following another vehicle, with k_{dlc} a factor (just) larger than one. The third term expresses that the space gap on the target lane $d_t(t)$ must be larger than a threshold d_{lc} . The lateral driving model for changing lane to the right can be formulated in the same way by formulating conditions for a vehicle to drive (almost) at the intended speed.

The intelligent function models all equipment that influences the driver behavior. Examples of intelligent functions are traffic lights, driver warning systems for downstream congestion or Adaptive Cruise Control that detects a predecessor and controls the speed of the vehicle taking into account a safe headway to the vehicle. The general operation of the intelligent functions follows three steps. First, data is collected from onboard sensors and by receiving data from entities outside the vehicle, such as other vehicles or traffic control systems at the road side. Second, the data is fused and used to assess the ambient conditions around the vehicle using state estimation and processing. The third step is to generate actions, such as providing information or a warning to a driver, to (partially) control the vehicle, or to transmit data to the outside world. The intelligent function is input to the driver model. It may replace elements of the driver model, for instance, in the case of Adaptive Cruise Control. It may also model the interaction of the system with the driver, for example, the response of a driver to a warning given by the systems resulting in a cautious acceleration behavior.

The basic description of traffic flow simulation models already allows for simulation of traffic on a simple motorway stretch. It may be extended to more realistic scenarios, including on-ramps, off-ramps, rural and urban roads, and intersections. Such extensions will also require route choice modeling of drivers.

4 Concept Development

4.1 Introduction

Concept development is the first stage of a product design or policy development process. It can start with nothing more than an idea for an ITS function or with the societal and/or individual purpose that this function should fulfill. It should lead to a more specific scoping of the ITS application to be designed, and in particular to the design requirements of this application. This means that concept development does not lead to a real prototype or application, but only to a prototype or application “on paper.” The problem at hand is to develop specific requirements on the ITS application, based on rather generic goals and conditions.

In policy development, simulation can be used to develop road maps for the deployment of ITS in conjunction with stated policy goals, in order to mitigate the risk of adopting ineffective or inefficient policies. The policy goals are typically formulated on the societal level, in terms of societal impacts such as traffic efficiency, traffic safety, and the environment.

In product design, simulation can be used to explore the potential of ITS functions in meeting stated business goals and conditions. The purpose of employing simulation is to mitigate the cost of prototyping by virtual exploration. Ultimately, the business goals and conditions are intrinsic to the company in question, but they are usually reformulated to cater to the individual customer or the society as a whole. Thus they may refer to the societal impacts mentioned above, as well as driver comfort and satisfaction, technical performance, HMI effectiveness, etc.

4.2 Cooperative Adaptive Cruise Control

The use of simulation in the development is illustrated for a concept for Cooperative Adaptive Cruise Control (CACC). Research has shown that Adaptive Cruise Control can improve traffic throughput but may lead to instable traffic (Pueboobpaphan and van Arem 2010). Cooperative Adaptive Cruise Control is an Adaptive Cruise Control that can communicate with other equipped vehicles. Simulation is used to assess to which extent CACC can improve traffic flow stability. More details may be found in Schakel et al. (2010).

The following assumptions are used for the operation of CACC, following the CACC2 controller in Wilmink et al. (2007), which is capable of operating in traffic flows of vehicles with and without CACC. This allows investigation into the effects at multiple penetration rates. An important building block in the controller is a regular Adaptive Cruise Control. The Adaptive Cruise Control acceleration of vehicle d is defined in (7.5).

$$a_d = \min(k_{cc}(v_{cc} - v_x), \quad k_2 e_v + k_1 e_x) \quad (7.5)$$

Here k_{cc} is a constant gain, v_{cc} is the desired speed, v_x is the vehicle velocity, e_v is the relative speed error (downstream vehicle speed minus v_x), e_x is the relative distance error, and k_1 and k_2 are gains. The CACC controller extends this control law by including the speed differences with additional leaders. Only equipped vehicles are included with a maximum distance of 200 m to the concerned vehicle. A maximum of $n = 5$ leaders is used in (7.6). As CACC also operates with mixed traffic, a relation to the distance headway is excluded.

$$a_d = \min \left(k_{cc}(v_{cc} - v_x), \quad k_2 e_{v;d-1} + k_1 e_{x;d-1} + \left(\frac{k_2}{n-1} \sum_{i=d-n}^{d-2} e_{v;i} \right) \right) \quad (7.6)$$

4.3 Simulation Setup

Human drivers and CACC-equipped vehicles were simulated on a 4 km stretch of road with a single lane. The first vehicle drives at 90 km/h for the first 80 s covering the first 2 km. Then the vehicle decelerates at a rate of -5 m/s^2 to a speed of 36 km/h. This speed is maintained for 5 s after which the vehicle accelerates at a rate of 1 m/s^2 back to 90 km/h. As this initial perturbation forms a deceleration to 36 km/h, there is room left for shockwave growth due to unstable behavior. A Gaussian distribution was applied on the desired time headway both for human drivers and CACC-equipped vehicles. This will introduce small headways that may result in unstable behavior but also large headways that may result in more stable behavior. By varying the headway standard deviation, the net effect can be assessed. Note that both human drivers and CACC-equipped vehicles have the same desired headway as the assessment concerns traffic flow stability and not capacity. Two headway distributions have been used, $1.2 \pm 0.15 \text{ s}$ and $1.2 \pm 0.3 \text{ s}$. The average of 1.2 s gives a capacity of 2,400 veh/h at a desired speed of 90 km/h. The distribution of headways introduces slight decelerations at the start of the stretch of road as the vehicles are generated at a fixed headway corresponding to the inflow. This slightly decreases inflow capacity. The inflow is set at 2,000 veh/h. Penetration levels of 0%, 50%, and 100% have been evaluated.

4.4 Results

Trajectories of each vehicle describing position and speed as a function of time were used to assess the shockwave dynamics. The following steps describe the derivation of the location and speed of a shockwave: First, for each vehicle, find the first five successive time steps with a deceleration stronger than -1 m/s^2 . The first of five time steps is the anchor point (x, t) of the shockwave. This leads to a set of anchor points. Second, find the least-squares solution with $x = f(t) = x_0 + v_s t$ through the anchor points. The shockwave speed is then given by V_s and x_0 is nothing more than a spatial intercept. Third, the stopping criterion is defined as a maximum allowable distance error depending on the shockwave

■ Table 7.1

Shockwave characteristics. (Based on Schakel et al. 2010, with permission)

CACC penetration rate	Headway 1.2 ± 0.3 s			Headway 1.2 ± 0.3 s		
	Duration (s)	Range (m)	Speed (m/s)	Duration (s)	Range (m)	Speed (m/s)
0%	Unstable	Unstable	-4.4	Unstable	Unstable	-4.4
50%	9.0	-173.6	-18.7	9.0	-171.1	-18.5
100%	7.6	-579.5	-76.6	7.9	-572.2	-73.6

speed. With larger shockwave speeds, larger distance errors are allowed. During each iteration, the distance errors are divided by the latest shockwave speed versus. This results in a “travel time” for which a maximum error of 8 s is the stopping criterion. The results are displayed in ● Table 7.1 and are based on an average of 10 stochastic runs. The shockwave duration is deduced from the maximum and minimum time of the anchor points of all remaining vehicles. The shockwave range (covered distance) is derived from the duration and shockwave speed. For the shockwave speed the last value of versus is used. If the stopping criterion of the shockwave is not reached within the duration of the simulation, the shockwave is characterized as unstable.

The results clearly show that CACC indeed leads to stable traffic flow, whereas manual traffic (based on IDM+) turns out to be unstable. An increase in penetration rate leads to a slightly shorter duration but a much larger range. Further, the results show that CACC leads to much higher shockwave speeds, especially when compared with manual traffic. Finally, the increase in headway variability appears to have a very limited effect on traffic flow stability.

4.5 Discussion

Although the simulation results have confirmed the ability of CACC to improve traffic flow stability, the resulting fast shockwaves may be undesirable, especially to human drivers in the case of mixed traffic flows. For all downstream CACC vehicle except the direct predecessor, the CACC controller is sensitive to speed differences only, and other aspects such as the position and acceleration are only indirectly of influence. It is may be expected that taking into account the positions of the preceding vehicles may reduce the shockwave speed, while maintaining the stabilizing impacts. For CACC systems that use acceleration differences, the value of k_2 influences the shockwave speed as this governs the acceleration response and hence acceleration differences. Lower values of k_2 and fewer vehicles that are anticipated for will result in slower shockwaves. Nonetheless, these shockwaves will be faster than without CACC as any CACC system should anticipate further ahead than humans ($n > n_{human}$ and $k_2 > 0$) to have more stability in traffic flow. Given the larger shockwave speeds of CACC, human drivers may be less able to anticipate

and either show more unstable behavior or increase headways leading to a decrease in capacity relative to a theoretical model capacity.

5 Application Testing and Verification

5.1 Introduction

Once the concept of an intelligent function is developed and demonstrated successfully, the intelligent function can be developed further into an application. Application development includes development of new technologies that enable the application of the intelligent function. WiFi communication is an example of such a new technology that enables the cooperation of ACC-equipped vehicles into CACC vehicles in the previous section, or the basic set of applications defined by ETSI (ETSI 2009). The application of the CACC involves the development of the software components to handle the communication, generate some form of situational awareness to keep track of the ordered string of leading vehicles and their motion state relative to the ego vehicle. These software components typically reside in the facility and application layers of the ETSI architecture (ETSI 2010). The systems for communication, positioning, and the application platform running the application software have to be developed and integrated into the vehicle. These software modules and support systems construct the application and realize the intelligent function of the initial concept.

Simulation plays an important role in the testing and verification of the technology and the application prior to field testing and demonstrations. This section describes simulation as part of the laboratory test environment for testing and verification of the components and systems at various stages of development and integration.

5.2 Requirements on Simulation in a Laboratory Test Environment

The main purpose of a laboratory test environment is to provide a controlled environment in which the behavior of the application can be tested during development and before the application has been integrated into vehicles for field testing. Typically, the test environment is centered around a single vehicle – the ego vehicle. The environment of the ego vehicle is simulated in a laboratory setup, and includes, for example, the simulation of neighboring traffic, equipped vehicles and roadside units, and communication on a local scale. The laboratory test environment serves several purposes:

- Debugging and testing of the isolated application components, that is, unit testing
- Debugging and testing for integration of the units into applications
- Stress testing of applications under controlled conditions
- Verification of application behavior against the specifications
- Verification and validation of the application under severe and dangerous conditions that cannot be realized in field tests

- Early evaluation of the concepts for flaws under more realistic conditions before field testing can commence

The test environment replicates relevant and critical conditions of the design envelope. In addition to the typical ideal environment used for concept development, it simulates, for example, (near) crashes, unsafe traffic situations, sensor and hardware failures or intermittent failures, or erroneous driver interactions. This obviously poses significant requirements on the fidelity of the models and simulators used.

The laboratory environment is used during development of the individual parts of the application, that is, unit testing, and the integration of these units into a complete application in the lab setup. For unit testing, a fully virtual environment is required that implements the interfaces of the unit and simulates the behavior of the environment of the unit, including the system behavior of neighboring units.

When units are coupled and integrated into partial systems, the simulation environment should have the flexibility to evolve as well. The simulation models or simulators previously neighboring the unit under testing are now replaced by the real components. Such a test environment is also called a test bench. The boundaries, interfaces, and behavior of the simulated environment have to be expanded. The lab setup should accommodate testing the hardware and software modules of the final system in so-called Software-in-the-Loop (SiL) and Hardware-in-the-Loop (HiL). The lab setup may also include a driving simulator for developing, testing, and verifying human-machine interactions. This poses significant requirements on the flexibility for configuring and interfacing the simulators; see, for example, PreDRIVE-C2X (2009a, b; 2010). Here an example from the SAFESPOT project is used, where a test bench was set up to support various stages of testing and integration of cooperative functions in the vehicles and on roadside units.

5.3 An Example: The SAFESPOT Test Bench

SAFESPOT was an integrated project in the Sixth Framework Program, running 2006–2010, for the development of cooperative vehicles and road infrastructure for road safety. SAFESPOT developed the concept of a Safety Margin Assistant (SMA). The SMA is a set of cooperative applications that extends drivers' awareness of the surrounding environment in space and time, detect potentially dangerous situations in advance, and prevent road accidents. Applications run both on vehicles and on roadside units. Typical applications are speed limitation and safety distance, collision warning, curve warning, road condition warning, vulnerable road user warning, and intersection safety.

The SAFESPOT platforms on the vehicles and roadside units are very similar. The platform on the vehicles, called SAFEPROBE, consists of a network with computers for positioning, a router for communication, vehicle sensors and a gateway to the CAN bus, data fusion, and SMA applications.

SAFESPOT is a large project in which many geographically distributed and multidisciplinary teams are developing parts of the system and different applications. Yet all have

to work together properly. Hence, a need for a common methodology to verification and validation as presented in De Gennaro et al. (2009), including a common set of test scenarios and a laboratory test environment is crucial. One of the test environments, the Simulator Test Bench (see Netten and Wedemeijer (2010)), is used here as an example.

The Test Bench consists of a SAFEPROBE platform and a submicroscopic traffic simulator called MARS (Fig. 7.2) and a monitoring tool (see De Gennaro et al. (2009), not included in Fig. 7.2). The Application PC (Fig. 7.2, right) runs an application platform with the SMA application software and application support software. The applications are decomposed into Driver Assistance Applications (DAA) and Cooperative Support Applications (CSA). The DAA form the heart of the cooperative safety functionality as it continuously monitors the environment of the ego vehicle for unsafe situations and reacts by informing the driver or taking automatic actions. The CSA contains the functionality to cooperate with neighboring vehicles and roadside units.

The SAFEPROBE platform (Fig. 7.2 middle right) runs facilities such as the Local Dynamic Map (LDM) and Data Fusion to provide a common situational awareness to all applications, Vehicle Adhoc NETWORKing (VANET) for WiFi communication, and provides gateways to the positioning and vehicle sensors.

The Simulator Test Bench is set up to test the platform and SMA at increasing levels of integration. The test bench uses the simulator for simulating the ego vehicle's systems and sensors, neighboring vehicles' behavior, local traffic scenarios, and the VANET communication. Figure 7.3 shows the internal structure of the simulated ego vehicle. Four consecutive setups are implemented with specific interface components between the simulation environment and the application.

5.3.1 Setup 1: Application Unit Testing

The test bench is used initially as a virtual test environment in which a SAFESPOT application component is developed and tested as a unit in isolation. The unit, implemented in C++, can be inserted into a controller of the ego vehicle in MARS. For example, the driver assistance application (DAA#1) is a unit to determine when an advice should be sent to the driver for the safety distance and speed advice application. The DAA#1 can run in parallel with a CACC controller in the simulator. The sensor input for headway distance and relative vehicle speeds can be extracted from the MARS vehicle, and the output should be forwarded to a driver model in the same MARS vehicle. The performance can be evaluated in the simulator in terms of the time-to-collision or headway at which the advice is generated by DAA#1.

5.3.2 Setup 2: Testing the Integration of Applications with Facilities

The next step of integration is to run a single application component on the Application PC and test the interface to an underlying facility such as the Local Dynamic Map (LDM).

Simulated data from the ego vehicle's world model is fed into the LDM. A MARS LDM Data Player (MARS-LDP) feeds the motion state data from an ego vehicle and other road users directly into the LDM. The MARS-LDP can run off-line in a step mode for debugging, in-line with a simulation, or replay data from field trials. The LDM is queried by the application component, and the real LDM and the application unit can be tested.

5.3.3 Setup 3: Application Platform Testing

The Application PC is now integrated with the SAFESPOT platform. The platform is used in the laboratory environment and the data from the sensors, vehicle gateway and VANET router is simulated. A MARS-UDP plug-in is developed for all vehicle input devices, that is, laser scanner, radar, VANET router, positioning PC, and vehicle gateways. A MARS-UDP plug-in is "plugged" in the respective sensor or router model of the ego vehicle in MARS and replicates the simulated sensor output as UDP messages to the SAFESPOT platform. The software modules from the real SAFESPOT components for constructing the UDP messages are reused in the plug-ins to guarantee that exactly the same UDP message formats are generated in the simulations.

Setup 3 is used to test the components and integration of the SAFESPOT platform. This setup is also used to test the integration of the Application PC and platform as well as the integration of application components.

5.3.4 Setup 4: Testing the Cooperation with External Vehicles and Roadside Units

Finally, the output from the applications is fed back into the simulation to close the simulation and test loop. This also closes the loop for cooperation, via communication and sensing, with neighboring vehicles and roadside units. This interface is realized by plugging in a MARS-UDP server in the VANET router of a MARS vehicle. This server receives the UDP messages from the SAFESPOT network and broadcasts the message via MARS to the intended receivers.

The SAFESPOT Test Bench shows how a simulation environment can be used to support the testing and integration of applications at various levels of integration, that is, from isolated software components, via SiL to HiL testing. Adaptations to the simulation environment are required to accommodate the expanding interfaces of the integrated application. The advantage of using a single simulation environment is that a common set of simulation scenarios, which are defined for testing and verification of the application specifications, can be reused by multiple partners and throughout the integration process. Such a joined approach and laboratory test environment can reduce integration issues and development efforts considerably.

6 Application Validation and Evaluation

This phase concerns the validation and evaluation of intelligent applications. It can arise either as the third phase in the application development process, after the Testing and Verification phase, or as an evaluation of existing applications, for example, prompted by policy makers or prospective buyers. Validation only arises in the first setting and involves checking that the prototype satisfies earlier postulated user needs. Evaluation is a more open assessment of the performance and societal impacts of the application. Rather than checking whether certain preset needs are fulfilled, it seeks to quantify the achievements of the application. As mentioned before, the distinction between Verification and Validation is that Verification tests that technical requirements are met (inward scope) whereas Validation is user oriented (outward scope). The distinction between Validation and Evaluation is that Validation checks preset requirements while Evaluation answers open research questions. For further discussion, the distinction between Validation and Evaluation is not very relevant and therefore it is limited to evaluation for simplicity.

Validation and Evaluation can involve various kinds of analyses and analysis tools. The focus of this text is restricted to the use of simulation tools to determine the impacts of intelligent applications.

The simulation setup used for Evaluation is rather similar to that for Concept Development. The main difference is that while Concept Development models conceptualized intelligent functions, Evaluation models concrete existing applications, with corresponding consequences for the interpretation of the outcomes.

The evaluation can cover both the technical performance and the societal impacts. For the first type, the simulation setup is very similar to that of [Sect. 5](#); therefore this chapter will focus on societal impacts. They are determined with traffic simulation, where models have to be included for the application, as well as the driver and vehicle interaction with the application. Deployment scenarios with varying penetration rates of the application (i.e., the fraction of vehicles that is equipped with the application), different application versions or settings and different environmental or traffic conditions can be simulated by changing parameters in the traffic simulator, with the outcomes feeding into an impact analysis. The models can be constructed using results from various sources such as verification tests, field operational tests or literature. The next subsection discusses a shockwave damping application as an example.

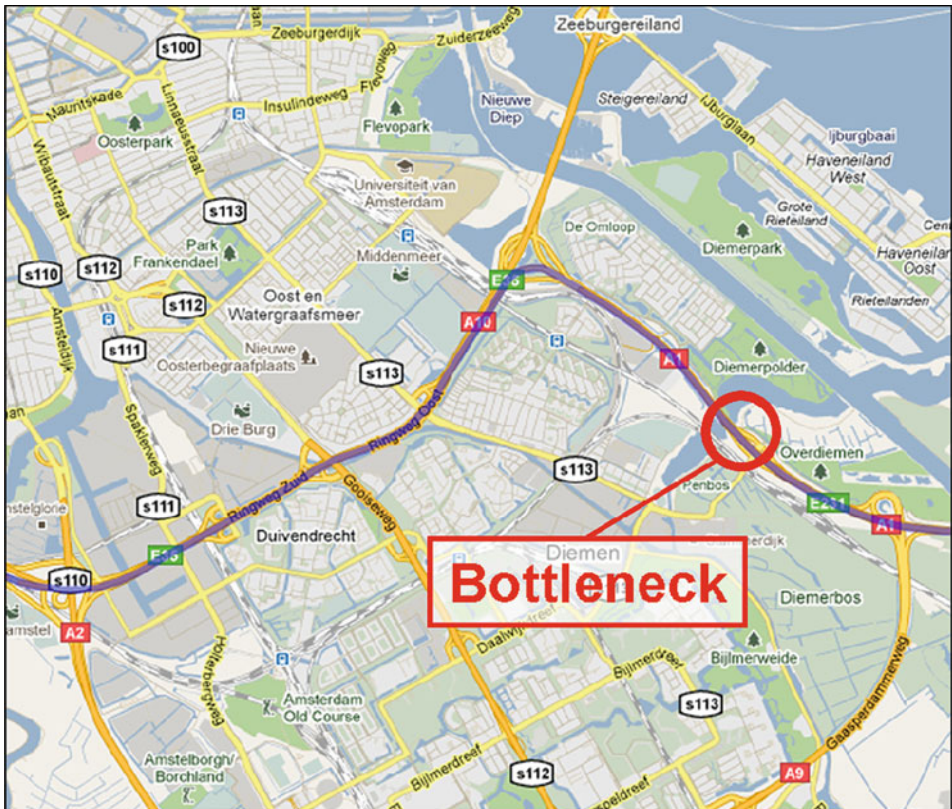
6.1 An Example: Shockwave Damping

In general, shockwaves are abrupt and large changes in the characteristics of traffic flow that travel upstream or downstream. The term is usually reserved for relatively small regions of congestion that travel upstream along an otherwise uncongested motorway, and is not (or no longer) attached to a bottleneck, and that is the meaning in which it will be used here. Shockwaves add up to a significant amount of congestion and form a safety hazard because they may appear suddenly and in unexpected locations. Preventing or

mitigating shockwaves will therefore significantly improve traffic flow. This section presents a simulation experiment with a cooperative application that aims to do this. This experiment is described in (Calvert et al. 2011). The text of this section summarizes this publication.

The application consists of an in-car system that communicates the vehicle motion state (position, speed, acceleration) to nearby vehicles via telecommunications, and controls the longitudinal motion of the vehicle in a way similar to the CACC described in ▶ Sect. 4.2. It also employs radar to detect the predecessor vehicle directly. The controller uses the motion state information of the predecessor vehicle to tailor its behavior to the prevention and mitigation of shockwaves.

Controlled field experiments have been conducted in 2010 (van den Broek et al. 2010), with encouraging results. The field experiments have naturally been limited in the sense that they involve a small number of vehicles on a short stretch of motorway without



■ Fig. 7.4

Simulated scenario. The main roads are shown in gray shade, some with route numbers; the simulated network consists of the east-bound lanes of the A10 and A1 motorways, overlaid in darker shade. At the bottleneck the number of lanes drops from 4 to 3. Figure reproduced with permission from (Calvert et al. 2011), copyright IEEE 2011

bottlenecks or ramps, the driver behavior is constrained by the experiment, and only the cases where all vehicles are equipped or all vehicles are unequipped have been tried. Naturally this leads to the question how the application will behave in real

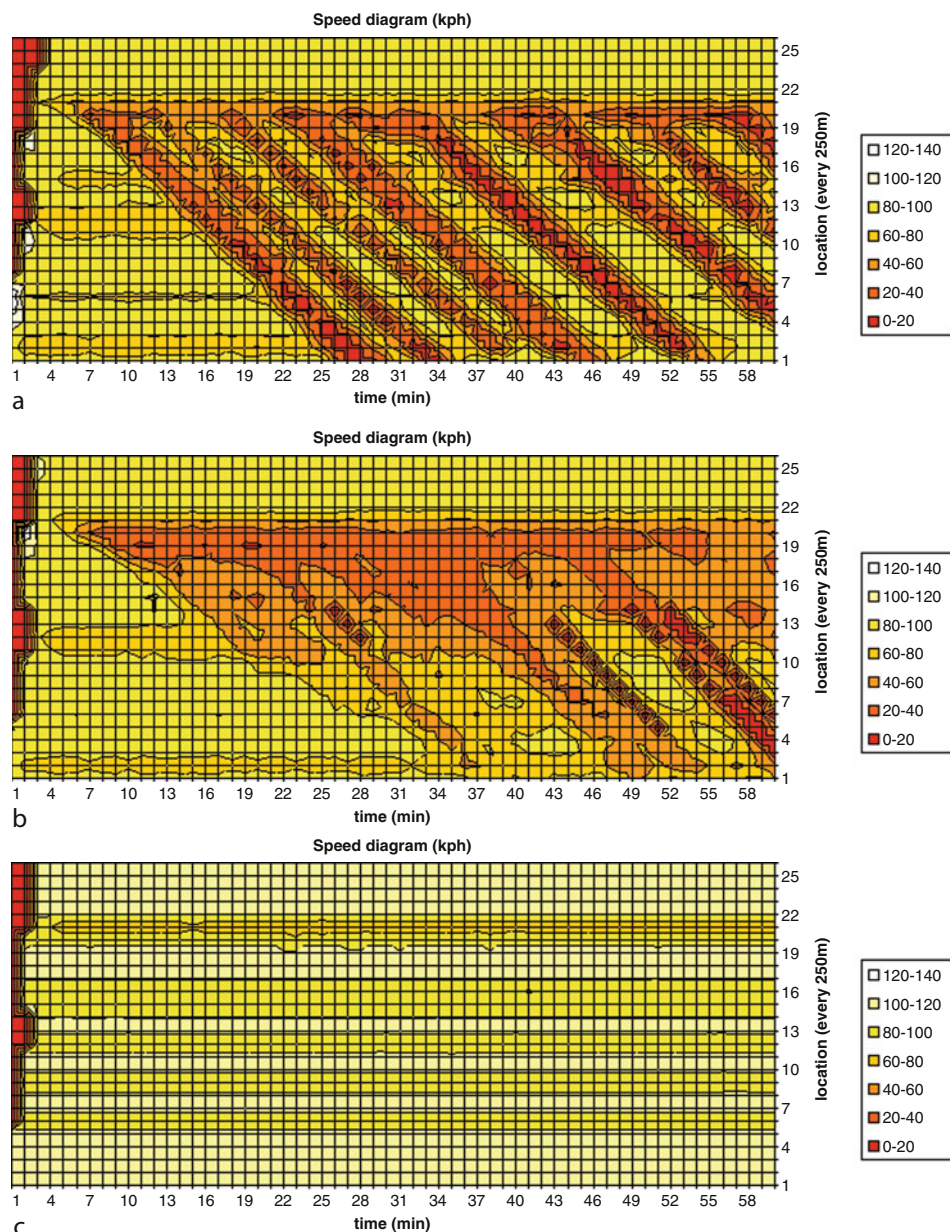


Fig. 7.5

(a, b, c) Speed diagrams for equipment rates 0% (top), 50% (middle), and 100% (bottom).

Figure reproduced with permission from (Calvert et al. 2011), copyright IEEE 2011

traffic. Experiments on a large scale are difficult and expensive to conduct in reality and therefore a simulation has been performed. The simulation scenario is a busy section of the motorway network in the Netherlands, consisting of the interconnected motorways A1 and A10 south of Amsterdam, see ● Fig. 7.4. The speed limit on these roads is 100 km/h. Daily peak traffic demand is high enough to generate shockwaves. In the scenario, these shockwaves emanate from the indicated bottleneck where the number of lanes drops from 4 to 3. The scenario is simulated with the microscopic traffic simulation tool *ITS Modeler* (Versteegt et al. 2005). This tool allows the experimenter to insert any desired vehicle-driver model, and this feature is used to import the shockwave damping application that has been used in the field experiment into the simulator.

The evaluation is set up as a comparison between scenarios where different percentages of vehicles are equipped with the shockwave damping application and a reference scenario where all vehicles are unequipped. The performance of the application is measured using the *average network speed* and the *cumulative arrival of vehicles at their destinations* as performance criteria. Furthermore space-time speed diagrams are surveyed to gain insight into the dynamics of the traffic flows. ● Figure 7.5 shows some of these diagrams. At 0% equipment rate, one clearly sees shockwaves originating around location 21 and traveling upstream. At 50% equipment rate, these have been dissolved into a larger region of less-severe congestion. At 100% equipment rate, the congestion has disappeared completely. The underlying data shows that the *average network speed* and the *cumulative arrival of vehicles at their destinations* increase monotonically with the equipment rate, and hence that the application has a positive effect even at low equipment rates.

7 Conclusions and Challenges

The development of systems for intelligent systems in vehicles is a complex process, involving many technological components, requirements with respect to robustness and fail safety, time pressure, and different stakeholders. In this chapter, the different roles of simulation have been shown in the development process from the early concept development, through testing and verification of components to the validation and evaluation of systems. Basic traffic flow simulation can be used during the concept development phase. Detailed simulation including simulation of the dynamic behavior of technical components can be used for testing and verification. Advanced traffic flow simulations can be used for validation and evaluation of applications tested in a Field Operational Test.

Consistency between simulation tools used in the different phases of the development of intelligent systems in vehicles is important to safeguard transferability of results between these different phases. This chapter has shown how these simulation tools can be combined into an integrated framework, allowing for consistent simulation at different levels of detail, varying from driver models to detailed sensor models, using different temporal and geographical scales. Moreover, the framework integrates the different disciplines needed in the development process, ranging from traffic engineering, systems engineering, human factors, communication technology, information

technology, and sensor technology. The framework enables advanced and efficient testing procedures using Software-in-the-Loop and Hardware-in-the-Loop simulation. Finally, the framework offers interfaces to other tools such as driving simulators.

The development of simulation tools is an ongoing process. This chapter has highlighted the integration of simulation tools in a single framework to provide better support to the development of intelligent systems in vehicles. Future challenges for simulation tools are the standardization of interfaces between simulation models, of simulation data, and transferability of model components. The use of open source software is expected to play an important role in this challenge. A second challenge concerns the further development, calibration, and validation of the mathematical functions that describe the dynamic behavior of the simulation objects. In particular, this concerns the development, calibration, and validation of models for driving behavior including the response of drivers to intelligent systems in the vehicle. Third, and finally, further development of the simulation tools is expected to move toward virtual or mixed reality using advanced visualization techniques

References

- Brackstone M, McDonald M (1999) Car-following – historic review. *Transp Res Pt F* 2:181–196
- Van den Broek THA, Netten BD, Hoedemaeker M, Ploeg J (2010) The experimental setup of a large field operational test for cooperative driving vehicles at the A270. In: *Proceedings of IEEE conference on intelligent transportation systems (ITSC)*, Madeira, pp 198–203
- Calvert SC, van den Broek THA, van Noort M (2011) Modelling cooperative driving in congestion shockwaves on a freeway network. In: *Proceedings of IEEE conference on intelligent transportation systems (ITSC)*, Washington, DC (to appear)
- De Gennaro MC et al (2009) Verification and validation of SAFESPOT vehicle based applications. In: *ITS 2009*, Stockholm
- ETSI (2009) Basic set of applications, European telecommunications standards institute ETSI. <http://www.etsi.org>, Technical Report EN TR 102 638, v1.1.1, 2009–06
- ETSI (2010) ITS communication architecture, European Telecommunications Standards Institute ETSI. <http://www.etsi.org>, Draft specification EN 302 665 v1.0.0 2010–03
- Gietelink O, Ploeg J, De Schutter B, Verhaegen M (2006) Development of advanced driver assistance systems with vehicle hardware-in-the-loop simulations. *Veh Syst Dyn* 44(7):569–590
- ISO (2004) Systems and software engineering – system life cycle processes. ISO/IEC 15288. <http://www.iso.org>
- Netten B, Wedemeijer H (2010) Testing cooperative systems with the MARS simulator. In: *ITSC2010*, Madeira
- van Noort M, van Arem B, Park B (2010) MOBYSIM: an integrated traffic simulation platform. In: *Proceedings of ITSC*, Madeira
- OSI (1994) Open systems interconnection – basic reference model. ISO/IEC 7498–1:1994. <http://www.iso.org>
- PreDRIVE-C2X (2009a) Description of user needs and requirements, and evaluation of existing tools. PreDRIVE-C2X Deliverable D2.1. <http://www.pre-drive-c2x.eu/>
- PreDRIVE-C2X (2009b) Test bench: description and specification of testing tools and software components. PreDRIVE-C2X deliverable D3.2. <http://www.pre-drive-c2x.eu/>
- PreDRIVE-C2X (2010) Description of communication, traffic and environmental models, and their integration and validation. PreDRIVE-C2X deliverable D2.3. <http://www.pre-drive-c2x.eu/>
- Pueboobpaphan R, van Arem B (2010) Driver and vehicle characteristics and platoon and traffic flow stability – understanding the relationship for design and assessment of cooperative adaptive

- cruise control. Transportation Research Record: Journal of the Transportation Research Board, No. 2189. Transportation Research Board of the National Academies, Washington, DC, pp 89–97
- Schakel WJ, van Arem B, Netten B (2010) Effects of cooperative adaptive cruise control on traffic flow stability. In: IEEE conference on intelligent transportation systems, proceedings, ITSC 2010, Madeira Island. Article number 5625133, pp 759–764
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. Phys Rev E 62:1805–1824
- US DOD (2009) Technology readiness assessment (TRA) deskbook, US Department of Defense. http://www.dod.mil/ddre/doc/DoD_TRA_July_2009_Read_Version.pdf
- Verhoeff L, Verburg DJ, Lupker HA, Kusters LJJ (2000) VEHL: a full-scale test methodology for intelligent transport systems and vehicles and subsystems. Proceedings of the IEEE intelligent vehicles symposium (IV), Detroit, October 2000, pp 369–375
- Versteegt E, Klunder G, van Arem B (2005) Modeling cooperative roadside and invehicle intelligent transport systems using the ITS Modeller. In: ITS World, San Francisco
- Wilmink IR, Klunder GA, van Arem B (2007) Traffic flow effects of integrated full-range speed assistance (IRSA). In: Proceedings of 2007 IEEE Intelligent vehicles symposium, Istanbul, pp 1204–1210

Section 2

Vehicle Longitudinal and Lateral Control Systems

Azim Eskandarian

8 Vehicle Longitudinal Control

Jihua Huang

Partners for Advanced Transportation Technologies (PATH),
Institute of Transportation Studies, University of California at
Berkeley, Richmond, CA, USA

1	<i>Introduction</i>	168
1.1	History of Vehicle Longitudinal Control	169
1.2	Functionality of Vehicle Longitudinal Control	170
2	<i>System Requirements and Framework Design</i>	171
2.1	System Requirements	171
2.2	System Framework	173
3	<i>Vehicle Longitudinal Control</i>	174
3.1	Sensors for Longitudinal Control	174
3.2	Longitudinal Vehicle Model	175
3.3	Longitudinal Control System Design	176
3.3.1	Upper Level Controller	177
3.3.2	Lower Level Controller	178
3.4	Experimental Results	179
4	<i>Integrated Lateral/Longitudinal Control</i>	180
5	<i>Longitudinal Control of Automated Buses</i>	184
5.1	Bus Precision-Stopping Control	184
6	<i>Conclusion</i>	187

Abstract: Decades of research efforts have greatly advanced our understanding of vehicle longitudinal control and yielded fruitful results. This chapter provides a detailed discussion on this integral part of vehicle regulation control. This chapter starts with an introduction that defines the scope of our discussion as the longitudinal control of automated vehicles, provides the relevant research history, and describes the functionality of vehicle longitudinal control. Subsequently, the system requirements and framework design are discussed. As a safety-critical system, the longitudinal control of automated vehicles needs to satisfy both safety requirements and performance requirements. Typically formulated as a feedback control system, the longitudinal control system consists of sensors (and sensor processing), control computation, and control actuation components. This chapter further describes the longitudinal control systems for passenger vehicles in detail, which covers the sensing, modeling, and controller design by using a longitudinal control system designed for automated vehicles in a platoon as an example. To further extend the discussion to automated heavy vehicles, a specific control application, precision-stopping control for automated buses, is discussed. This chapter concludes with a short summary of the current status and thoughts on future directions for the longitudinal control of automated vehicles.

1 Introduction

Vehicle control technologies can substantially improve the handling and safety of road vehicles. Depending on the automation level, vehicle control functions can be divided into three categories:

- Active safety systems, which include antilock braking system (ABS), traction control, stability enhancement systems, rollover control, and collision warning/avoidance systems
- Semi-automated vehicle functions, i.e., driver assistance systems, which include adaptive cruise control (ACC) and active front steering
- Fully automated vehicle control, which replaces the driver in conducting driving functions

In active safety systems, the driver is still in control of the vehicle. The systems are automatically activated (assuming the driver does not turn the function off) only under emergency situations where the vehicle might go out of control. These systems also try to achieve the desired acceleration/deceleration and trajectory indicated by the driver's control input. The semi-automated vehicle, on the other hand, can have full control of specific vehicle functions. For example, the ACC can control the longitudinal motion of the vehicle without the driver's interference. Usually, such systems must be manually activated by the driver, and the driver can retake control over the vehicle at any instant. These semi-automated functions free the driver from specific driving controls, but they usually do not have capabilities beyond their specific functions. Limited control of the vehicle and a lack of decision-making capabilities are the major differences between

semi-automated vehicle functions and fully automated vehicle control. Nonetheless, the specific control functions can serve as basic servo controls for the latter; e.g., the ACC is indeed closely related to the automated longitudinal control.

The fully automated vehicle control represents the highest automation among the three categories. It can completely replace the driver in all vehicle functions and tasks. It not only has full control of vehicle longitudinal and lateral motions, but also has the capability to determine necessary maneuvers and can even decide which route to take to arrive at its destination. In this chapter, the longitudinal control of fully automated vehicles is the main focus, while vehicle lateral control and ACC as well as CACC are the topics of another two chapters, respectively.

1.1 History of Vehicle Longitudinal Control

The history of advanced vehicle control systems (AVCS), which includes the longitudinal control of automated vehicles, was described in detail in (Shladover 1995). This history suggests that the automated longitudinal control as well as the development of automated vehicles is closely related to that of the automated highway systems (AHS). The AHS was first shown as a feature of the General Motors “Futurama” exhibit of future transportation opportunities at the 1939–1940 New York World Fair. During the late 1950s, General Motors and RCA developed and demonstrated (on test tracks) automatic control of the steering and speed of automobiles for what they called the “Electronic Highway” (Gardels 1960; Zworykin and Flory 1958). From 1965 to 1980, a long-term research program at Ohio State University (OSU) on both steering and longitudinal (spacing) control (Fenton and Mayhan 1991) was conducted under the sponsorship of the Ohio Department of Transportation and the Federal Highway Administration (FHWA).

Interest in automated vehicles was revived as the “modern” history of AVCS development began in 1986 with parallel actions in Europe and the USA. In Europe, the PROMETHEUS Program (Program for European Traffic with Highest Efficiency and Unprecedented Safety) was initiated by the motor vehicle industry, with funding support from the European Community’s EUREKA program and the governments of the major western European Countries. The program was intended to provide a framework for cooperative development to improve the capacity and safety of road transportation systems. In 1986, the California Department of Transportation (Caltrans) joined with the Institute of Transportation Studies of the University of California at Berkeley to found the PATH (Partners of Advanced Transportation Technologies, originally called Program on Advanced Transit and Highways) Program. The thrust areas of the PATH program include clean propulsion technologies, highway automation, and road vehicle navigation and guidance (Shladover et al. 1991).

In Japan, fully automated driving has been sponsored by the Ministry of International Trade and Industry (MITI) via the Personal Vehicle System (PVS) and Super-Smart Vehicle Systems (SSVS) (Tsugawa et al. 1991). Meanwhile, the Ministry of Construction, as part of its Advanced Road Transportation System (ARTS) project, has been researching

on the fully automated highway concept, including automation of freight movements in exclusive goods distribution tunnels located deep beneath major cities in Japan (Fujioka et al. 1993).

Longitudinal control of automated vehicles has received attention since the 1960s, and possibly even earlier. A detailed review was conducted by Shladover (1995), in which the longitudinal control strategies were classified into 12 structures according to the sources of the feedback information of the control. Among them, the feedback of spacing and velocity difference relative to a preceding vehicle is a popular controller structure because of its simplicity and its potential for use in mixed traffic (Recently, this feedback control structure has become extremely popular for use in ACC systems). Longitudinal control of a platoon, on the other hand, feeds back information communicated from other vehicles in the platoon, as well as the relative spacing and speed information. For example, Tsugawa (Tsugawa and Murata 1990) uses the feedback of the complete state of the preceding vehicle and controlled vehicle, plus the commanded velocity of the preceding vehicle. The PATH longitudinal controllers, on the other hand, feedback the speed and acceleration of the platoon leader, as well as the complete state of the controlled vehicle and its predecessor.

1.2 Functionality of Vehicle Longitudinal Control

Vehicle longitudinal control together with vehicle lateral control forms the regulation control of an automated vehicle, which controls the vehicle's longitudinal and lateral motions via feedback laws (algorithms) to implement the desired maneuvers. While vehicle lateral control is a function of each individual vehicle relative to the roadway reference, longitudinal control adds the complication of potential interactions among multiple controlled vehicles. Historically, three configurations were explored to handle these interactions or to avoid them: fixed block control (Pitts 1972), point following (or moving block) control (Wilkie 1972; Fenton and Mayhan 1991), and vehicle following control (Rajamani et al. 2000; Fenton and Mayhan 1991). Among them, vehicle following control systems provide essentially all of the longitudinal regulation control functionality for the vehicle, with none required to be on the roadside. Due to its potential to significantly increase lane capacity and reduce the computational costs, vehicle following control has become the main stream since 1980s.

In vehicle following control configuration, the control applied to a vehicle, in a line or a platoon, is determined by its longitudinal state with respect to the other vehicles. In the simplest case, i.e., autonomous longitudinal control, the control is dependent only on the state with respect to the nearest lead vehicle; while in the most complex case, i.e., longitudinal control of a platoon of vehicles, the control may depend on its states with respect to all vehicles in the platoon. Consequently, the control of a platoon demands considerably more communication capability than that of autonomous longitudinal control. The vehicle-following studies at Ohio State University from 1964 to 1971 mainly involve autonomous longitudinal control, and focus on basic problems associated with vehicle longitudinal control. The longitudinal control at California PATH, on the other

hand, focuses more on platooning. The motivations for this change include both the maturity of computing and communication technologies and the substantially greater lane capacity offered by the platooning approach.

Whether a vehicle is under autonomous longitudinal control or operates in a platoon, the most basic function of the longitudinal control is to control the speed of the host vehicle and to maintain its distance from the preceding vehicle. In addition, longitudinal control functions also include intentional changes in vehicle separation to accommodate maneuvers such as joining or splitting from platoons and merging of traffic streams. For the latter functions, coordination between the longitudinal control and the lateral control (and possibly the coordination with nearby vehicles via communications) is involved (Interested readers can read the control of splitting and merging processes in (Rajamani and Shladover 2001; Rajamani et al. 2000) for more details.)

This chapter is organized as follows. 🔗 Section 8.2 describes system requirements and introduces the framework design for vehicle longitudinal control. 🔗 Section 8.3 describes the longitudinal control systems for passenger vehicles in detail, which covers the sensing, modeling, and controller design. 🔗 Section 8.4 extends the discussion to automated heavy vehicles with a specific control application: precision-stopping control for automated buses. 🔗 Section 8.5, summarizes the current status and discusses future directions for the longitudinal control of automated vehicles.

2 System Requirements and Framework Design

2.1 System Requirements

From a broad viewpoint, vehicle longitudinal control as part of the regulation control of automated vehicles need to satisfy the two key requirements for automated vehicles: safety and performance. Safety indicates reliability and robustness; performance highlights the consistency of the vehicle and passenger comfort. These requirements are also true when the system works under abnormal conditions such as adverse environment and system failures.

Safety issues of automated vehicles can come from the environment and the automatic control system. Environmental factors include road/weather conditions and interaction between vehicles sharing the same road, while safety factors involve hazards, faults, and system failures. To ensure vehicle/road safety, the automated vehicle longitudinal control (as well as the automated vehicle lateral control) must possess the following capabilities:

- **Robustness:** The vehicle needs to maintain a bottom-line performance even under adverse road and weather conditions. If a satisfactory performance cannot be met, the system needs to be able to detect the problem, warn the driver, deactivate automatic functions, and switch to manual control.
- **Fault detection and management:** Faulty operation of the system hardware can degrade system performance or even induce complete system failure. Fault detection

and management allows the vehicle to detect faults, determine their impact on the system capabilities, and decide on appropriate reaction, such as executing degraded-mode control strategies or stopping the vehicle.

For fully automated vehicles, the longitudinal control should also have the capability to detect and avoid collision and to handle emergency situations:

- Collision detection and avoidance capability: Unless operating in dedicated and enclosed lanes, automated vehicles require the ability to detect obstacles on the road and to cope with surrounding vehicles in order to avoid collisions.
- Emergency response: Emergency response is the last resort in a deadly situation. A rapid and safe reaction to an emergency command input is therefore imperative.

The specific performance requirements for vehicle longitudinal control may vary, depending on applications and operating conditions. For example, the requirements for automated passenger vehicles in normal highway operations can be different from that for the precision stopping of automated buses. General performance requirements, however, usually include the following:

- Accuracy: The achievement of small deviation from a vehicle's desired state; e.g., its desired spacing and relative speed with respect to the preceding vehicle. Generally, the accuracy is expected to be higher than that achieved by average drivers.
- Ride comfort: Bounded acceleration/deceleration and jerk for smoothness; no noticeable oscillation and sufficient damping for passenger comfort. Ride comfort may be sacrificed only under emergency conditions when vehicle and occupant safety consideration may preclude comfort.
- Consistency: Consistent operation over the range of the expected environmental conditions and for all expected disturbance inputs.
- Smooth manual/automatic transition: Typically, the system allows both the automatic mode and the manual control mode (or even the semiautomatic mode); therefore, easy and smooth transition is critical to driver acceptance.
- Efficiency: Effective utilization of a vehicle's capabilities (such as acceleration capabilities) and efficient fuel usage.
- Cost effectiveness: Balance between the leverage of advanced technologies and practical cost and implementation concerns.

Typically, tracking accuracy and ride quality are considered as the primary performance requirements for vehicle longitudinal control. Tracking accuracy is important for maintaining safety and for enabling vehicles to operate in close proximity to each other so that they can increase lane capacity and reduce aerodynamic drag (thereby also saving fuel and emissions). Ride quality is important for driver and passenger acceptance but is also closely related to the ability of the control system to save fuel and to reduce pollutant emissions.

Another performance requirement critical to vehicle longitudinal control is the so-called string stability (Swaroop and Hedrick 1996). In other words, string stability of an

interconnected system implies uniform boundedness of the state of all the systems. For example, in automated vehicle-following applications, tracking (spacing) errors should not amplify downstream from vehicle to vehicle for safety.

The string stability requirement results in different spacing strategies for the autonomous longitudinal control and that of a platoon. The autonomous systems are able to maintain string stability by adopting a constant time gap following control law, under which the separation between vehicles is proportional to their speed. The platoon can maintain string stability under this type of control law, but also under constant-spacing vehicle following, when the separation is invariant with speed (Swaroop et al. 1994). This latter mode makes it possible for electronically coupled platoons to operate at close spacing, even at high speeds, providing the opportunity to increase highway capacity significantly.

2.2 System Framework

The longitudinal control system can be divided using traditional block diagram schematics of feedback control systems. The main design components include the sensors (and sensor signal processing), control computation, and control actuation.

Sensing: Vehicle position and motion sensors play a vital role in the success of a vehicle's longitudinal control system, for the accuracy and reliability of the sensor measurements directly determine the system performance. While vehicle motion sensors usually include speed sensors and inertia navigation sensors (INS), such as accelerometers and gyros, various sensors are available to provide vehicle longitudinal position information. These sensors can be classified into two groups: (1) autonomous sensors, such as radar/LIDAR and cameras, which provide vehicle longitudinal positions relative to neighboring vehicles, and (2) cooperative sensing technologies that provide measurements of positions in the earth inertia coordinates. The latter usually employ Global Positioning System (GPS), which is typically integrated with the INS for improved accuracy, reliability, and bandwidth. This type of sensing needs the positions of neighboring vehicles (via inter-vehicle communication) in order to support the longitudinal control.

Each kind of sensors has its own strengths and weaknesses. While selecting sensors, one should take into account the various operational factors of the sensor, such as accuracy, repeatability, reliability, size, dynamic range, sensitivity, process requirement, and cost. In addition, taking sensor redundancy into account is critical to ensure the safety of automated vehicles. For example, the duplication of sensor sets results in redundancy; however, there might be a significant benefit in combining information from multiple sensors since different kinds of sensors have the potential to enhance system performance by compensating for each other in terms of their advantages and disadvantages.

Actuating: For vehicle longitudinal control, the control inputs include the throttle position and brake torque (or other equivalent inputs). Actuators receive the control command computed by the controllers and “actuate” the corresponding vehicle

subsystems so that the desired force can be delivered to the vehicle. The primary factors in actuator design include: (1) whether “dual use” is desirable; (2) the performance requirements for the actuators, such as the desired bandwidth and accuracy; and (3) servo control of actuators.

If “dual use” is preferred, the actuator should not interfere with manual operation. In such cases, the actuator is likely to be an add-on actuator, operating on top of the original manual actuating system. On the other hand, if “dual use” is not a concern, a new actuator might replace the original manual actuating system completely for the benefit of higher bandwidth and better control accuracy.

Usually, an actuator’s bandwidth should be at least five times that of the controlled system. If, for practical reasons (such as cost and implementation simplicity, as well as the existing vehicle resonance modes), the actuating system in vehicle control does not have a relatively high bandwidth, its magnitude and phase characteristics can have a significant impact on the control system design. In such cases, a servo controller needs to be designed to reduce the imposed restrictions.

Control Algorithm: The control algorithms involve both the servo control that improves the performance of actuators and the vehicle longitudinal control that achieves desired longitudinal motions. The next section will discuss the longitudinal control of a passenger vehicle in detail.

3 Vehicle Longitudinal Control

3.1 Sensors for Longitudinal Control

Four types of information are usually considered for the longitudinal control: (1) the speed and acceleration of the host vehicle; (2) the distance to the preceding vehicle; (3) the speed and acceleration of the preceding vehicle; and (4) the acceleration and speed of the first vehicle (i.e., lead vehicle) in the platoon. For vehicles operating in a platoon, knowledge of the acceleration and speed of the first vehicle (i.e., lead vehicle) in the platoon is often desired.

The speed and acceleration of the host vehicle can be measured by speed sensors and accelerometers onboard the vehicle. The distance to the preceding vehicle is usually measured by ranging sensors, such as radar, LIDAR (light detection and ranging), machine vision systems, and ultrasonic sensors. So far, radar has been the most commonly recommended sensor modality (Meinel 1995; Grimes and Jones 1974; Alvisi et al. 1991). LIDAR’s performance is often affected by adverse weather, particularly in fog and snow conditions; and also needs a lens hole at the front of the vehicle. Machine vision systems face the challenges of dealing with significant variation in environmental conditions (bright sunlight, darkness, fog, rain, or snow) and recovering 3D data. Ultrasonic transducers’ range limitation severely restricts their role as vehicle sensors.

Theoretically, the speed and acceleration of the preceding vehicle can be derived from that of the host vehicle and the measurements (i.e., range and range rate) from the

range sensor. However, the estimation of the relative acceleration requires differentiation of the measurements from the range sensor, which can be noisy. Another approach is to use communications to transmit the speed and acceleration of a vehicle to the vehicle that follows it. The platoon concept facilitates this approach. Similarly, the speed and acceleration of the lead vehicle of a platoon can also be transmitted to vehicles in the platoon to help realize string stability. However, the reliability of communication becomes a concern in this approach. Consequently, autonomous longitudinal control generally uses the first two types of information, as well as the speed of the preceding vehicle, while the longitudinal control of platooning takes advantage of all four types of information.

3.2 Longitudinal Vehicle Model

The automotive power train was partitioned into the following segments: an engine, a transmission (including a torque converter), a drivetrain (including rubber tires), and any other components that can influence the longitudinal performance of an automobile (such as throttle, fuel control, spark control, EGR system, clutches, bands, brakes, accessories, etc.). A 12-state nonlinear longitudinal model (Hedrick et al. 1993) was developed based upon the previous developed longitudinal models of a typical front wheel drive vehicle equipped with a V-6 engine (Cho and Hedrick 1989; Moskwa and Hedrick 1989). The model includes four states for the engine, two for the transmission, and six for the drivetrain, as well as two time delays associated with the engine. Due to the complexity of the model, it is primarily used for simulation.

For the control design, the 12-state complex model is further simplified based on the following assumptions (Hedrick et al. 1993): (1) time delays associated with power generation in the engine are negligible; (2) the torque converter in the vehicle is locked; (3) there is no torsion of the drive axle; and (4) slip between the tires and the road is zero. These assumptions relate the vehicle speed v_x directly to the engine speed ω_e :

$$\dot{x} = v_x = Rh\omega_e \quad (8.1)$$

where R and h are gear ratio and tire radius, respectively. The mode is then reduced to a three-state nonlinear model (8.2, 8.3, and 8.6) with mass of air in the intake manifold (m_a), engine speed (ω_e), and brake torque (T_{br}) as the states (Rajamani et al. 2000).

The dynamics relating engine speed ω_e to the pseudo-inputs, net combustion torque T_{net} and brake torque T_{br} , can be then modeled by:

$$\dot{\omega}_e = \frac{T_{net} - c_a R^2 h^2 \omega_e^2 - R(hF_f + T_{br})}{J_e} \quad (8.2)$$

where c_a is the aerodynamic drag coefficient, F_f is the rolling resistance of the tires, and $J_e = I_e + (mh^2 + I_\omega)R^2$ is the effective inertia reflected on the engine side.

$T_{net}(\omega_e, m_a)$ is a nonlinear function obtained from steady-state engine maps available from the vehicle manufacturer. The dynamics relating m_a , the air mass flow in engine manifold, to the throttle angle can be modeled as:

$$\dot{m}_a = \dot{m}_{ai} - \dot{m}_{ao} \quad (8.3)$$

where m_{ai} and m_{ao} are the flow rate into the intake manifold and out from the manifold, respectively. m_{ao} is a nonlinear function of ω_e and P_m , pressure of the air in engine manifold (from the engine manufacturer); and

$$\dot{m}_{ai} = MAXTC(\alpha)PRI(m_a) \quad (8.4)$$

where MAX is a constant dependent on the size of the throttle body, $TC(\alpha)$ is a nonlinear invertible function of the throttle angle, and PRI is the pressure influence function that describes the choked flow relationship which occurs through the throttle valve. The ideal gas law is assumed to hold in the intake manifold:

$$P_m V_m = m_a R_g T \quad (8.5)$$

where V_m is the intake manifold volume, R_g is a variable that depends on the vehicle transmission gear ratio, and T is the temperature. Finally, the brake model is linear and modeled by a first-order lag:

$$\tau_{br} \dot{T}_{br} + T_{br} = T_{br,cmd} = K_{br} P_{br} \quad (8.6)$$

where T_{br} is the brake system time constant, K_{br} is the total proportionality between the brake line pressure P_{br} and the brake torque at the wheels, and the pressure distribution is assumed to be evenly split between the front and rear tires' brake torque constant of proportionality.

3.3 Longitudinal Control System Design

Since researches conducted at California PATH have significantly advanced the status of vehicle longitudinal control, the PATH controller is used here as an example for the design of longitudinal control. As the longitudinal control of a platoon, the PATH longitudinal control feedbacks information that is communicated from other vehicles in the platoon, as well as the relative spacing and speed information. More specifically, the PATH longitudinal controller feedbacks the speed and acceleration of the platoon leader, as well as the complete state of the controlled vehicle and its predecessor. This feedback structure has been applied to the three-state nonlinear vehicle longitudinal model described in the previous section (8.2, 8.3, and 8.6 with the mass of air in the intake manifold (m_a), engine speed (ω_e), and brake torque (T_{br}) as the states) (Shladover 1991; Rajamani et al. 2000), as well as models that include explicit representations of power-train dynamics (Hedrick et al. 1993). The controller design approaches have included linear pole-placement (Shladover 1991) and nonlinear sliding mode control (Hedrick et al. 1993; Rajamani et al. 2000). This section summarizes the longitudinal

controller PATH demonstrated in the 1997 NAHSC Demo (Rajamani et al. 2000), which was based on a modification of the standard sliding surface control technique.

This longitudinal controller consists of two levels of control: the upper level controller determines the desired or “synthetic” acceleration (there are also lots of designs where the upper level controller determines the desired velocity, instead of acceleration (Kato et al. 2002) for each vehicle in the platoon; the lower level controller determines the throttle and/or brake commands required to track the desired acceleration.

3.3.1 Upper Level Controller

The upper level controller determines the desired acceleration for each vehicle so as to (1) maintain constant small spacing between the vehicles; (2) ensure string stability of the platoon (see explanation below). For the upper level controller, the plant model is

$$\ddot{x}_i = u_i \quad (8.7)$$

where the subscript i denotes the i th vehicle in the platoon; x_i is its longitudinal position. The acceleration of the vehicle is thus assumed to be the control input u_i . However, due to the finite bandwidth associated with the lower level controller, each vehicle is actually expected to track its desired acceleration imperfectly. The performance specification on the upper level controller is therefore to meet objectives (1) and (2) robustly in the presence of a first-order lag in the lower level controller performance:

$$\ddot{x}_i = \frac{1}{\tau s + 1} \ddot{x}_{i,des} = \frac{1}{\tau s + 1} u_i \quad (8.8)$$

Equation 8.7 is thus assumed to be the nominal plant model while the actual plant model was given by Eq. 8.8.

The spacing error for the i th vehicle is defined as $\epsilon_i = x_i - x_{i-1} + L$, where ϵ_i is the longitudinal spacing error of the i th vehicle, with L being the desired spacing. In terms of spacing error, the above objectives of the upper controller can be mathematically stated as follows:

$$\epsilon_{i-1} \rightarrow 0 \Rightarrow \epsilon_i \rightarrow 0 \quad (8.9)$$

$$\|H(s)\|_{inf} \leq 1 \quad (8.10)$$

where $\hat{H}(s)$ is the transfer function relating the spacing errors of consecutive vehicles in the platoon: $\hat{H}(s) = \epsilon_i / \epsilon_{i-1}$.

The string stability of the platoon (objective (2), Eq. 8.10) refers to a property in which spacing errors are guaranteed to diminish as they propagate toward the tail of the platoon [40]. For example, string stability ensures that any errors in spacing between the second and third cars do not amplify into an extremely large spacing error between cars 7 and 8 further down in the platoon. In addition to Eq. 8.10, a condition that the impulse response function $h(t)$ corresponding to $\hat{H}(s)$ does not change sign is desirable (Swaroop and Hedrick 1996; Swaroop et al. 1994). The reader is referred to (Swaroop and Hedrick 1996) for details.

Longitudinal control algorithms that guarantee string stability in the platoon of vehicles include autonomous, semiautonomous, and radio communication-based algorithms. A comparison of the performance and deployment advantages of the three types of algorithms is provided in Rajamani (Rajamani and Zhu 2002). The NAHSC demonstration implemented a radio communication-based algorithm described below. The sliding surface method of controller design (Slotine and Li 1991) is used. Define the following sliding surface:

$$S_i = \dot{\epsilon}_i + \frac{\omega_n}{\xi + \sqrt{\xi^2 - 1}} \frac{1}{1 - C_1} \epsilon_i + \frac{C_1}{1 - C_1} (v_i - v_l) \quad (8.11)$$

Setting

$$\dot{S}_i = -\lambda S_i, \text{ with } \lambda = \omega_n \left(\xi + \sqrt{\xi^2 - 1} \right) \quad (8.12)$$

The desired acceleration of the vehicle is then given by

$$\begin{aligned} \ddot{x}_{i_{des}} &= (1 - C_1) \ddot{x}_{i-1} + C_1 \ddot{x}_l - 2 \left(2\xi - C_1 \left(\xi + \sqrt{\xi^2 - 1} \right) \right) \\ &\quad \omega_n \dot{\epsilon}_i - \left(\xi + \sqrt{\xi^2 - 1} \right) \omega_n C_1 (v_i - v_l) - \omega_n^2 \epsilon_i \end{aligned} \quad (8.13)$$

The control gains to be tuned are C_1 , ξ , and ω_n . The gain C_1 takes on values $0 \leq C_1 \leq 1$ and can be viewed as a weighting of the lead vehicle's speed and acceleration. The gain ξ can be viewed as the damping ratio and can be set to one for critical damping. The gain ω_n is the bandwidth of the controller. Equation 8.12 ensures that the sliding surface converges to zero. If all the cars in the platoon use this control law, results in (Swaroop and Hedrick 1996; Swaroop et al. 1994) show that the vehicles in the platoon are able to track the preceding vehicle with a constant spacing and that the system is string stable, i.e., the spacing errors never amplify down the platoon. Results on the robustness of the above controller, especially to lags induced by the performance of the lower level controller, can also be found in (Swaroop and Hedrick 1996). A wireless radio communication system is used between the vehicles to obtain access to all of the required signals. Each vehicle thus obtains communicated speed and acceleration information from two other vehicles in the platoon: its preceding vehicle and the lead vehicle. Setting $C_1 = 0$ for a two-car platoon, we obtain the following classical second-order system:

$$\ddot{x}_{i_{des}} = \ddot{x}_{i-1} - 2\xi \omega_n \dot{\epsilon}_i - \omega_n^2 \epsilon_i \quad (8.14)$$

3.3.2 Lower Level Controller

In the lower level controller, the throttle and brake actuator inputs are determined so as to track the desired acceleration described in Eq. 8.13. The controller was developed by

applying a modification of the standard sliding surface control technique (Slotine and Li 1991) to the simplified three-state model (8.2, 8.3, and 8.6). If the net combustion torque T_{net} is chosen as:

$$(T_{net})_i = \frac{J_e}{Rh} \ddot{x}_{ides} + [c_a R^3 h^3 \omega_e^2 + R(hF_f + T_{br})]_j \quad (8.15)$$

then, from (8.2), the acceleration of the vehicle equals the desired acceleration defined by the upper level controller: $\ddot{x}_i = \ddot{x}_{ides}$.

Since the pressure of air in the manifold P_m and temperature T can be measured, m_a is then calculated from the ideal gas law (8.5). With the required combustion torque from (8.15), the map $T_{net}(\omega_e, m_a)$ is inverted to obtain the desired air mass flow in engine manifold m_{ades} . A single surface controller is then used to calculate the throttle angle α to make m_a track m_{ades} . Define the surface as:

$$s_2 = m_a - m_{ades} \quad (8.16)$$

Setting $\dot{s}_2 = -\eta_2 s_2$ we have:

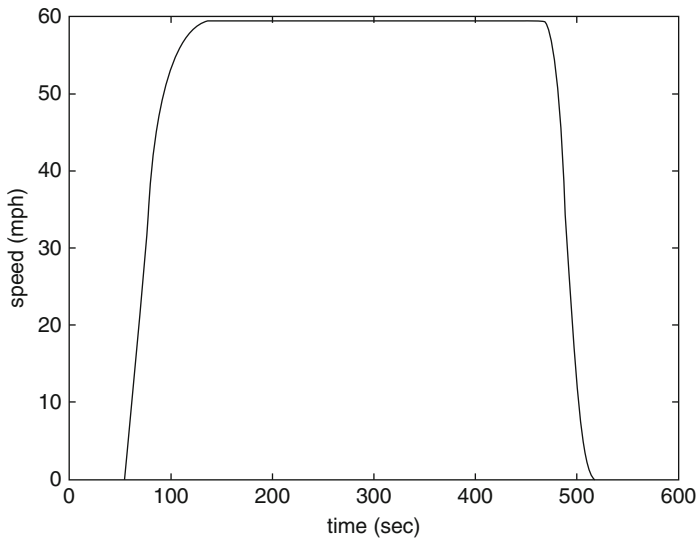
$$MAXTC(\alpha)PRI(m_a) = \dot{m}_{ao} - \dot{m}_{ades} - \eta_2 s_2 \quad (8.17)$$

Since $TC(\alpha)$ is invertible, the desired throttle angle can be calculated from (8.17). If the desired net torque from (8.15) is negative, the brake actuator is used to provide the desired torque. An algorithm for smooth switching between the throttle and brake actuators (Choi and Devlin 1995) is incorporated into the longitudinal control system.

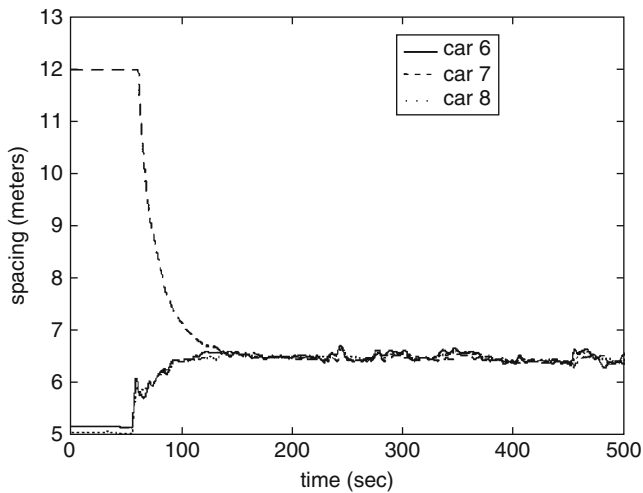
3.4 Experimental Results

The following figures document the performance of the PATH longitudinal controller during the 1997 NAHSC Demonstration. 8.1 shows the desired speed trajectory for the lead vehicle of an eight-car platoon. It starts from zero speed, ramps up, and then smooths out exponentially before cruising at approximately 60 mph. The deceleration profile consists of a ramp combined with a sinusoid. In the acceleration and deceleration trajectories, two gear changes (1–2 and 2–3) take place in the automatic transmission system.

8.2–8.4 show the spacing performances of cars 6, 7, and 8 which form the tail of the eight-car platoon. The cars start with arbitrary initial spacing. The steady-state desired spacing between the vehicles is 6.5 m. During the entire 7.6-mile run on the San Diego highway, the spacing error between these tail vehicles of the platoon remains within 0.2 m. This includes the spacing performance while the lead vehicle accelerates, cruises, decelerates to a complete stop, and other vehicles accelerate and decelerate while splitting and joining. The scenario also includes steep uphill and downhill grades during which the maximum spacing error occurs. 8.2 shows the initial arbitrary spacing between vehicles during start-up and the convergence to the steady-state 6.5-m spacing. 8.3 and 8.4 show close-ups of the inter-vehicle spacing to indicate the accuracy and smoothness of the longitudinal controller.



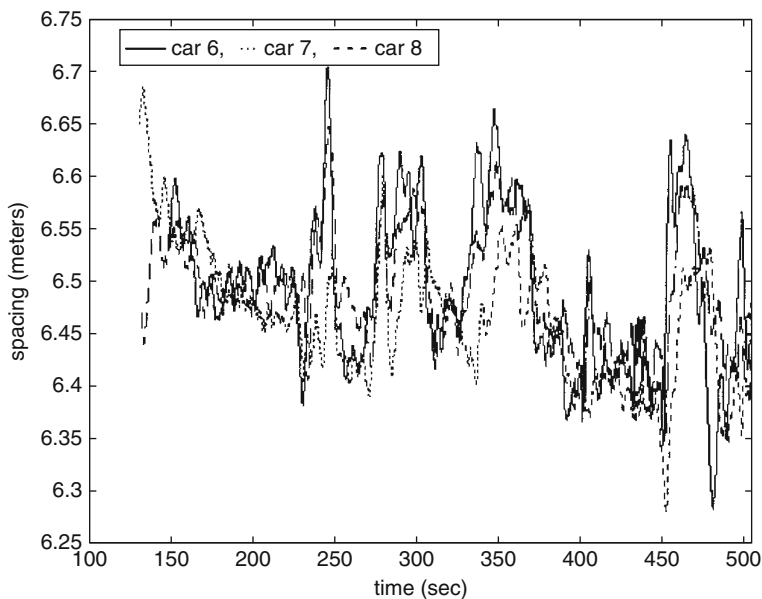
■ Fig. 8.1
Speed profile for the lead vehicle



■ Fig. 8.2
Spacing performance of cars 6, 7, and 8 of the eight-car platoon: raw radar data

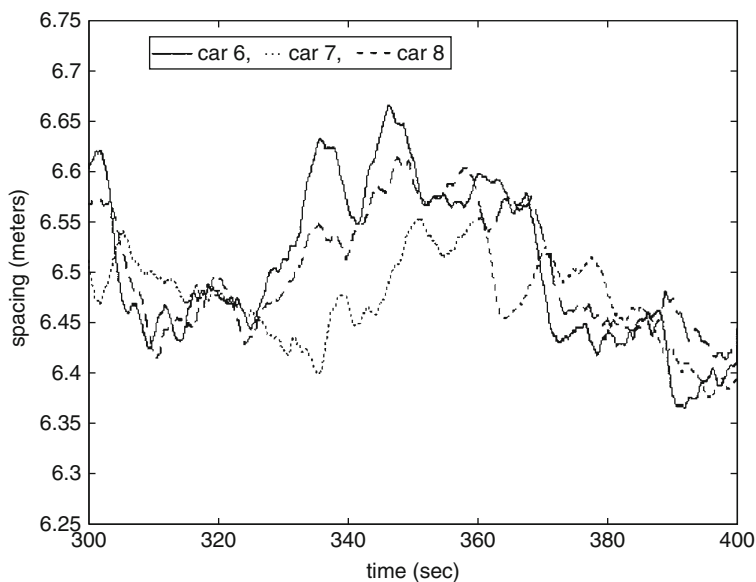
4 Integrated Lateral/Longitudinal Control

Although this chapter mainly focuses on vehicle longitudinal control, it is beneficial to extend the scope a bit wider by briefly introducing the coordination control that integrates longitudinal and lateral controls to carry out specific maneuvers. While the basic lane



■ Fig. 8.3

Spacing performance of cars 6, 7, and 8 of the eight-car platoon: estimated inter-vehicle spacing



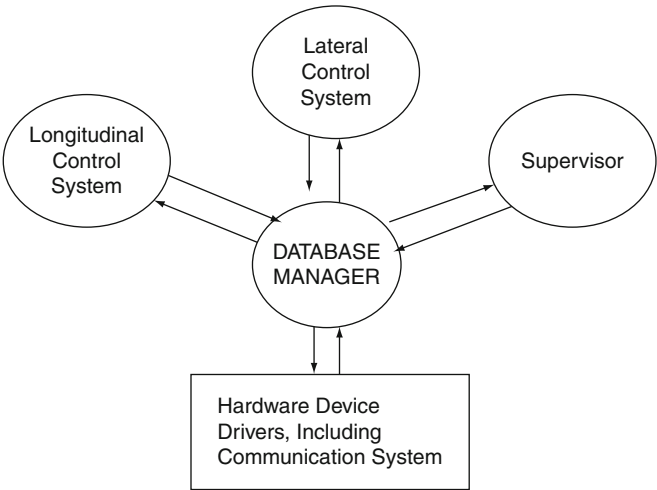
■ Fig. 8.4

Spacing performance of cars 6, 7, and 8 of the eight-car platoon: estimated inter-car spacing (blow-up plot)

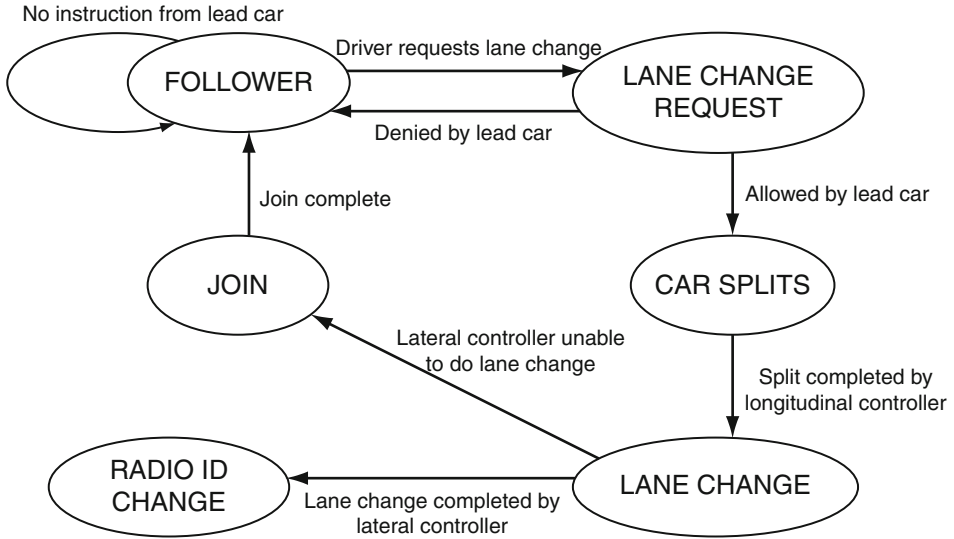
keeping and longitudinal spacing control can be performed independently, coordination between the two controllers is required to perform maneuvers like entry into a platoon, exit from a platoon, etc. A coordinated response is also required in the event of a sensor or actuator malfunction where a fault handling action has to be taken. ➤ [Figure 8.5](#) shows a structure of the integrated real-time software, which consists of four processes being executed simultaneously, with different sampling time and priorities assigned to each process. The four processes: longitudinal control system, lateral control system, device drivers, and supervisor, communicate with each other through the memory locations maintained by the database manager.

The supervisor shown in ➤ [Fig. 8.5](#) is an important process that coordinates the maneuvers to be performed by the lateral and longitudinal control systems. It retrieves sensor information and the status of the maneuver from the lateral and longitudinal systems as well as inputs from the driver of the vehicle through the database. Based on these inputs, the supervisor prescribes a desired maneuver for both the lateral and longitudinal controllers using finite-state machines.

➤ [Figures 8.6](#) and ➤ [8.7](#) provide an exemplary process of a lane change and the coordination actions performed by the supervisor. ➤ [Figure 8.6](#) shows the state machine for a vehicle when the driver has requested an exit from the platoon while being a follower. The request is transmitted to the lead vehicle which grants permission only if all the other vehicles in the platoon are in the FOLLOWER state and have not requested an exit. While granting permission to the vehicle to exit, the lead vehicle also instructs the vehicle behind the requesting vehicle to SPLIT (➤ [Fig. 8.7](#)) to a safe distance. When both the requesting vehicle and the vehicle behind have completed splitting, the lead vehicle grants permission to do the lane change. If the requesting vehicle is unable to

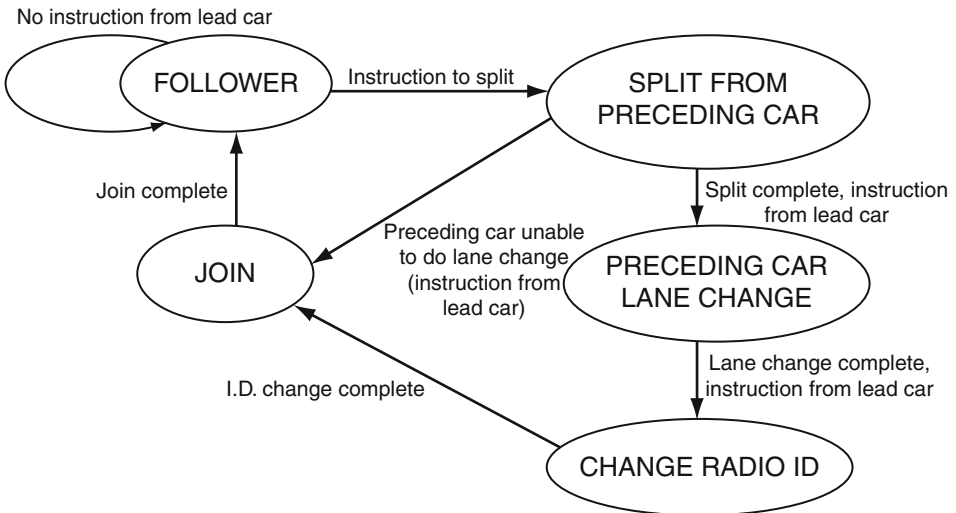


■ Fig. 8.5
Structure of the integrated real-time software



■ Fig. 8.6

Finite-state machine for the supervisor of the requesting vehicle



■ Fig. 8.7

Finite-state machine for the supervisor of the vehicle behind the requesting vehicle

do a lane change, it informs the lead vehicle and then moves to the JOIN state. After the JOIN state is completed, the vehicle comes back to the FOLLOWER state. In this case, the lead vehicle instructs the vehicle behind also to move to the JOIN and then FOLLOWER states.

5 Longitudinal Control of Automated Buses

The most likely application of the automated buses is automated Bus Rapid Transit (BRT), which has the potential to make the bus on time and allow more buses to run simultaneously to increase throughput. The automation functions in BRT could include precision docking, lane keeping, automated speed and space control, and maintenance yard operations. Among them, precision docking is a key automation function unique to automated buses. It enables buses to stop immediately adjacent to a loading platform, providing passengers with quick and easy boarding and alighting, even for those whose mobility is impaired. This subsection introduces the longitudinal control of the precision docking, i.e., bus precision stopping, as detailed in (Bu and Tan 2007). Its lateral control counterpart is described in (Tan et al. 2002).

5.1 Bus Precision-Stopping Control

Controlling a vehicle to stop smoothly and precisely with a consistency equal to or greater than those of an experienced operator is one of the longitudinal vehicle control functions. Other than bus precision docking, application examples also include backing automated trucks and trailers onto a platform, fueling automated trucks or buses, as well as stopping automatically at intersections. Since the majority of research on vehicle longitudinal controls focuses on the areas of high-speed platooning (Tsugawa and Murata 1990; Yanakiev and Kanellakopoulos 2001; Rajamani et al. 2000), adaptive cruise control (Liang and Peng 1999), and string stability (Swaroop and Hedrick 1996), this section describes the precision-stopping control of automated buses (Bu and Tan 2007) to extend the scope of our discussion on automated longitudinal control.

Two different approaches can be used to formulate the precision-stopping problem. One approach is the trajectory following. A desired trajectory is synthesized according to both the initial vehicle speed and position, and the final stop position. The controller is designed so that the vehicle will follow the desired trajectory with an appropriate brake command. The second approach is to dynamically synthesize a desired deceleration based on the vehicle's speed and remaining distance to the designated stop location. Brake servo command is generated to follow the desired deceleration. Between the two approaches, the trajectory-following approach is more intuitive and direct. By adopting this approach, the control problem can be formulated as follows (Bu and Tan 2007):

Given an initial vehicle speed v_0 , synthesize a brake control command u such that the vehicle follows the synthesized deceleration profile $x_{ld}(t)$ and stops at the desired location $x_{ld}(T)$ with a maximum error e_{max} and with desired smoothness represented by bounded deceleration a_{max} and jerk j_{max} .

Most buses and trucks today are equipped with pneumatic brake systems that use compressed air as the energy medium. Since the ability to remain “dual use” is one of the common requirement preferences for the early automation deployment requirements, it is desirable to use the existing pneumatic brake system as the primary means of stopping

control so that the automated vehicle can still maintain all its manual braking capabilities. Hence, the automatic brake control system can be realized by tabbing into the braking control commands of the pneumatic brake system or including an add-on actuator as described in (Bu and Tan 2007).

The dynamics of the pneumatic brake and the add-on brake actuator, together with vehicle longitudinal braking dynamics, are modeled and simplified to a three-state nonlinear model. A fifth-order polynomial trajectory is synthesized for the smooth stop. The challenges in the controller design are as follows:

- The pneumatic brake system is a highly nonlinear system because of the nonlinear pressure/airflow relationship.
- The pneumatic brake system, when coupled with heavy-duty vehicle longitudinal dynamics, has large uncertainties (vehicle load variation and parameter variation due to brake wear, temperature increase, and changes in road surface conditions).
- The system has unmatched model uncertainties since model uncertainties appear in equations that do not contain the control input.
- The vehicle longitudinal position is calculated by combining the vehicle velocity and the vehicle position information from magnet markers or transponders buried in the road surface, cameras with specific stripes on the road, or GPS receivers. However, information from magnet markers (which have approximately a 1-m interval in our setup) or transponders is often discrete and the GPS signal may be blocked by architecture around the bus station. Furthermore, position dead reckoning with vehicle speed may not work at low speeds since most vehicle velocity sensors can only sense a velocity that is larger than a certain speed (e.g., 0.6 m/s for CNG buses). This means that the longitudinal velocity and position information for many precision-stopping control systems may not be available or accurate enough during the final phase of vehicle stopping when accuracy is most needed.

To address these design difficulties, the following strategies are adopted. First, a physical model-based nonlinear analysis and synthesis is employed to address the nonlinear nature of the pneumatic brake system. Second, parameter adaptation is adopted to reduce the effect of modeling uncertainties. Specifically, the Indirect Adaptive Robust Control (IARC) approach (Yao 2003) is used to handle the general effects of model uncertainties. Third, the integrator backstepping design (Krstic et al. 1995) via Lyapunov function is used to address the mismatched model uncertainties. Finally, the accurate parameter estimation from the IARC parameters estimation is used to calculate the open-loop control command when longitudinal position information is not available at the final phase of vehicle stopping.

The designed precision-stopping control system was implemented on a 40-ft CNG bus for the Bus Precision Docking public demonstrations at Washington, D.C. in 2003. The demo bus is started manually by the driver. The driver can select the manual or automatic transition to the automatic control mode anytime he chooses. Once switched to the automatic control mode, the bus will automatically slow down or speed up to a predetermined cruising speed. When the bus reaches the location which is 12 m from

the designated final stopping point, it starts the precision-stopping process and stops exactly at the predetermined position along the station.

For over 50 total demonstration runs under various passenger load and road conditions, the final stopping accuracy was consistently controlled under 15 cm with the desired stopping smoothness, without a single failure. [Fig. 8.8](#) shows the tracking error for four different scenarios till the longitudinal position information becomes unavailable just before the final stop. (The scenario without parameter adaptation in the control algorithm is included for comparison.) Final stopping errors are measured manually when the bus is fully stopped. For the three scenarios with parameter adaptation, they are well kept within a 15 cm accuracy bound; for the case without parameter estimation, the final stopping error is larger than 30 cm.

For completeness, [Fig. 8.9](#) shows the lateral positions, steering angles, and speeds with respect to the magnetic marker numbers (1 m apart) during the D.C. Demo. It shows the high repeatability of the docking performance. The blow-up plots ([Fig. 8.10](#)) clearly illustrate that the bus has never touched the station, and the maximum error after the bus is approaching the station is within 1.5 cm peak-to-peak for front and 1 cm peak-to-peak for rear. The docking accuracy is about 1.5 cm peak-to-peak for all runs. Such high

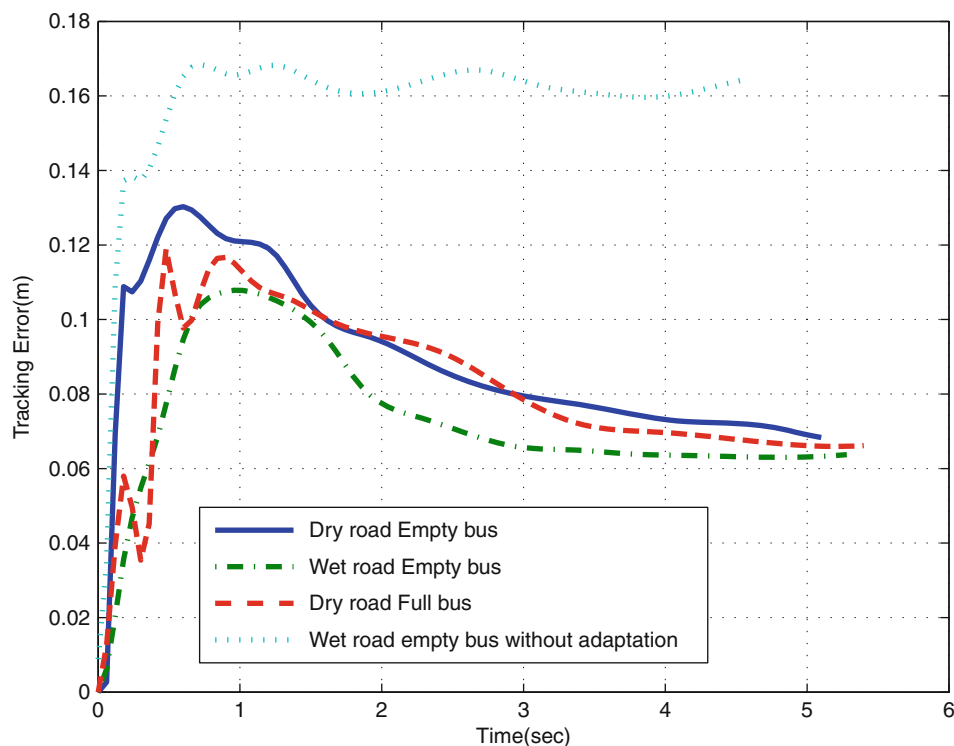
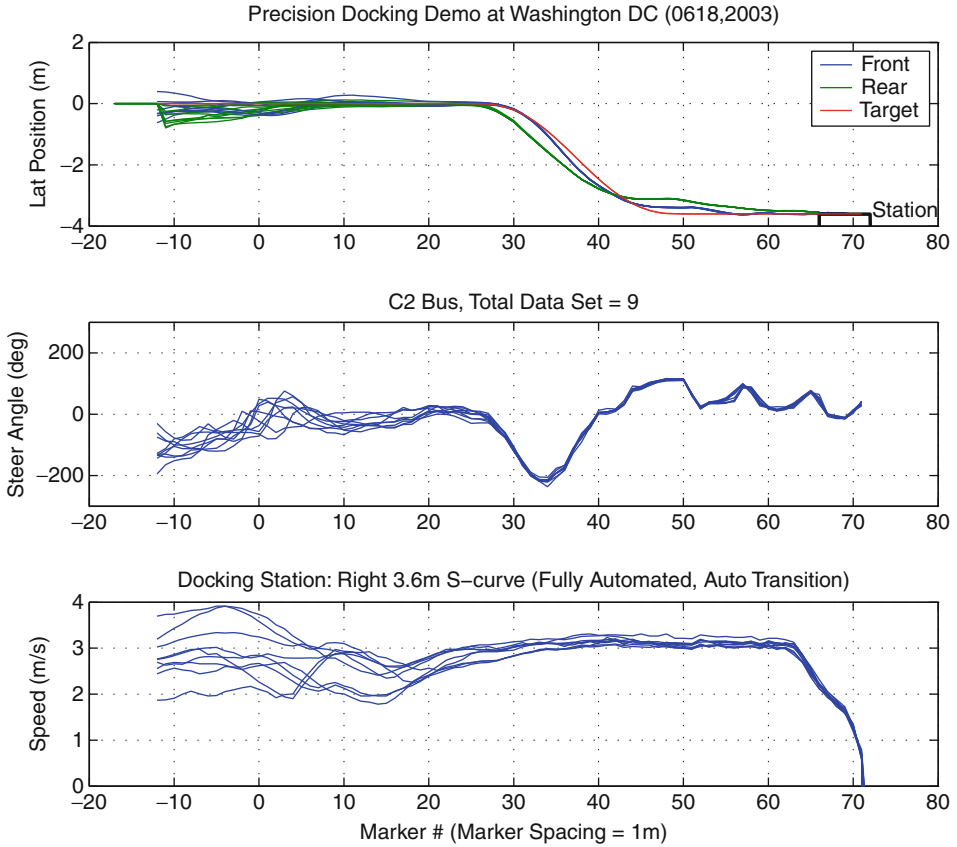


Fig. 8.8
Tracking error of the precision-stopping control for four different scenarios



■ Fig. 8.9

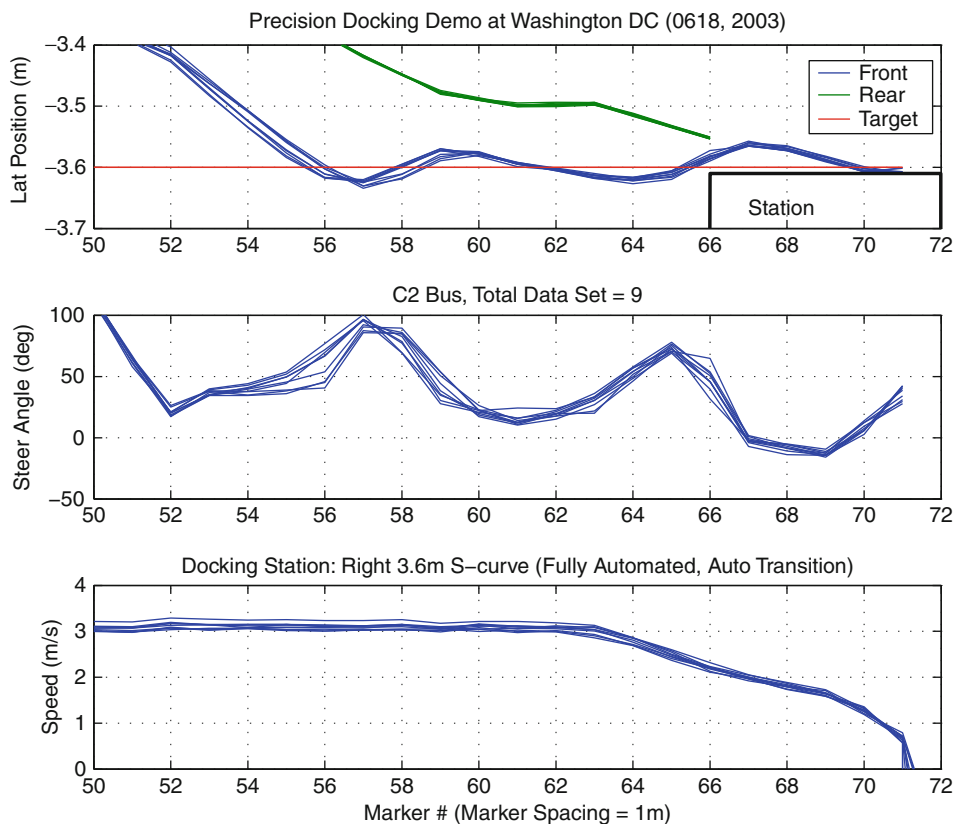
The lateral positions, steering angles, and speeds of bus precision docking in 2003 D.C. Demo

docking accuracies would allow fast loading and unloading of passengers similar to that of trains and greatly reduce the stress of manual docking in a high throughput Advanced Bus Rapid Transit system.

6 Conclusion

In this chapter, the longitudinal control of automated vehicles are described in detail, covering from the system safety and performance requirements, system configuration, longitudinal control design, to application examples such as an eight-car platoon operating on highways and bus precision stopping. The integration of the longitudinal control and the lateral control of an automated vehicle is also discussed briefly.

Decades of research efforts have greatly advanced our understanding of the longitudinal control of automated vehicles and yielded fruitful results. Various demonstrations of



■ Fig. 8.10

Blow-up plots of the lateral positions, steering angles, and speeds of bus precision docking in 2003 D.C. Demo

AHS systems and automated trucks and buses have demonstrated that the automatic longitudinal control system is capable of achieving a better and safer performance than a good driver. However, despite the great advances, safety-critical performances of vehicle longitudinal control have not been convincingly achieved or demonstrated. On one hand, those demonstrations are often conducted either on a test track or under a somewhat controlled environment, with vehicles operating only in a limited time span. Examples include AHS demonstration in closed highways, automated buses in dedicated lanes, and even the DARPA's Grand Challenge Program and Urban Challenge Program. On the other hand, commercial cruise control and adaptive cruise control typically operates in limited speed range and expect drivers to actively monitor the longitudinal performance and the driving environment in order to take over the control whenever necessary.

Moreover, most control-related researches in these areas focus on applications of control synthesis. A spectrum of control techniques has been explored and applied, including nonlinear optimization, back-stepping, and sliding mode and adaptive controls. So far, the majority of researches confine themselves to controller designs that

achieve adequate performance in normal (or benign) conditions. The difficulties inherent in real-world scenarios include vast amounts of uncertainties, short response time, broad operating range, and a large number of potential vehicles and obstacles. Although the longitudinal control has been researched for decades, there are still fundamental engineering questions to be answered. For example, a driver often switches between closed-loop control and open-loop (lightly closed-loop) driving and a typical emergency maneuver often requires a very quick switch between the two. Then what is the proper control architecture that captures the essence of such control? If it is a hierarchical control structure, what are the core “servo” control laws that can make the vehicle follow any supervisory commands as long as such commands are within its physical limits? What are the “worst-case” scenarios for longitudinal controls and how should such problems be modeled and dealt with in a reliable manner?

Furthermore, as an integral component of an automated vehicle, the longitudinal control will need to support the safety-critical performance of the integrated automated system. How should the longitudinal control and the integrated control system be modeled so as to reliably detect faults on an automated vehicle when there are so many uncertainties surrounding the vehicle? What is the control structure that can efficiently cope with various redundancies in sensors, computation, and actuation under strong cost constraints? How should layers of supervisory controls flexible enough to adapt to 10-plus years of continuous technology advancement be integrated into the system? And finally, how should safety and reliability be defined from a control and system standpoint so that the “liability” issues can have a more concrete “vehicle” to build upon?

The above questions can potentially redefine the tasks and responsibilities of control engineers in a future transportation system where the core of such system is an automatic control system.

References

-
- | | |
|--|---|
| <p>Alvisi M, Deloof P, Linss W, Preti G, Rolland A (1991) Anticollision radar: state of the art. Advanced telematics in road transport, vol 2. In: Proceedings of the DRIVE conference, Brussels, 1991</p> <p>Bu F, Tan H-S (2007) Pneumatic brake control for precision stopping of heavy-duty vehicles. <i>IEEE Trans Control Syst Technol</i> 15(1):53–64</p> <p>Cho D, Hedrick JK (1989) Automotive powertrain modeling for control. <i>ASME Dyn Syst Meas Control</i> 111(4):568–576</p> <p>Choi SB, Devlin P (1995) Throttle and brake combined control for intelligent vehicle highway systems, SAE 951897</p> <p>Fenton RE, Mayhan RJ (1991) Automated Highway Studies at the Ohio State University - an Overview. <i>IEEE Trans Veh Technol</i> 40(1):100–113</p> | <p>Fujioka T, Yoshimoto K, Takaba S (1993) A case study on an automated driving highway system in Japan. In: Transportation research board 72nd annual meeting, Washington, D.C.</p> <p>Gardels K (1960) Automatic car controls for electronic highways. In: General motors research laboratories report GMR-276</p> <p>Grimes DM, Jones TO (1974) Automotive radar: a brief review. <i>Proc IEEE</i> 62(6):804–822</p> <p>Hedrick JK, McMahon DH, Swaroop D (1993) Vehicle modeling and control for automated highway systems, PATH technical report, UCB-ITS-PRR-93-24</p> <p>Kato S, Tsugawa S, Tokuda K, Matsui T, Fujii H (2002) Vehicle control algorithms for cooperative driving with automated vehicles and intervehicle</p> |
|--|---|

- communications. *IEEE Trans Intel Transp Syst* 3(3):155–161
- Krstic M, Kanellakopoulos I, Kokotovic PV (1995) *Nonlinear and adaptive control design*. Wiley, New York
- Liang C, Peng H (1999) Optimal adaptive cruise control with guaranteed string stability. *Veh Syst Dyn* 32(4):313–330
- Meinel HH (1995) Commercial applications of millimeterwaves: history, present status, and future trends. *IEEE Trans Microw Theory Tech* 43(7):1639–1653
- Moskwa JJ, Hedrick JK (1989) Modeling and validation of automotive engines for control algorithm development. *J. Dyn. Syst. Meas. Control* 114 (2):278–286
- Pitts GL (1972) Augmented block guidance for short-haul transportation systems, applied physics laboratory report: APL JHU CP 019/TRP 023, John Hopkins University, p 178
- Rajamani R, Shladover SE (2001) An experimental comparative study of autonomous and cooperative vehicle-follower control systems. *Transp Res C Emerg Tech* 9(1):15–31
- Rajamani R, Tan H-S, Law BK, Zhang W-B (2000) Demonstration of integrated longitudinal and lateral control for the operation of automated vehicles in platoons. *IEEE Trans Contr Syst Technol* 8(4):695–708
- Rajamani R, Zhu C (2002) Semi-autonomous adaptive cruise control systems. *IEEE Trans Veh Technol* 51(5):1186–1192
- Shladover SE (1991) Longitudinal control of automotive vehicles in close-formation platoons. *ASME Dyn Syst Meas Control* 113:231–241
- Shladover SE et al (1991) Automated vehicle control development in the PATH program. *IEEE Trans Veh Tech* 40(1):114–130
- Shladover SE (1995) Review of the state of development of advanced vehicle control systems (AVCS). *Veh Syst Dyn* 24:551–595
- Slotine JJE, Li W (1991) *Applied nonlinear control*. Prentice-Hall, Englewood Cliffs
- Swaroop D, Hedrick JK (1996) String stability of interconnected systems. *IEEE Trans Autom Control* 41(3):349–357
- Swaroop D, Hedrick JK, Chien CC, Ioannou P (1994) A comparison of spacing and headway control laws for automatically controlled vehicles. *Veh Syst Dyn* 23(8):597–625
- Tan H-S, Bougler B, Zhang W-B (2002) Automatic steering based on roadway, markers: from highway driving to precision docking. *Veh Syst Dyn* 37(5):315–338
- Tsugawa S, Murata S (1990) Velocity control for vehicle following through vehicle/vehicle communication. In: *Proceedings of 22nd international symposium on automotive technology and automation*, Florence, Italy, pp 343–350
- Tsugawa S, Watanabe N, Fujii H (1991) Super smart vehicle system – its concept and preliminary works. *Proceedings of vehicle navigation and information systems symposium*, Dearborn, pp 269–277
- Wilkie D (1972) Moving cell motion scheme for automated transportation systems. *Trans Sci* 11(1):347–364
- Yanakiev D, Kanellakopoulos I (2001) Longitudinal control of automated chvs with significant actuator delays. *IEEE Trans Veh Technol* 50(5):1289–1297
- Yao B (2003) Integrated direct/indirect adaptive robust control of SISO nonlinear systems in semi-strict feedback form. In: *Proceedings of the American control conference*, Denver, Colorado, pp 3020–3025
- Zworykin VK, Flory LE (1958) Electronic control of motor vehicles on the highways. In: *Highway research board proceedings*, 37th annual meeting, Washington, USA

9 Adaptive and Cooperative Cruise Control

Fanping Bu · Ching-Yao Chan

California PATH, Institute of Transportation Studies, University of California at Berkeley Bldg., Richmond Field Station, Richmond, CA, USA

1	<i>Introduction</i>	192
2	<i>System Architecture and Operation Modes</i>	193
2.1	Overview and Operation Modes	194
2.2	System Architecture and Components	195
2.2.1	ACC/CACC Controller	196
2.2.2	Range Sensor	196
2.2.3	Vehicle States Sensing	196
2.2.4	Human–Machine Interface	196
2.2.5	Wireless Communication	198
3	<i>ACC/CACC Controller Design</i>	198
3.1	Control Problem Formulation and Design Objectives	198
3.1.1	Control Problem Formulation	198
3.1.2	Control System Structure	199
3.1.3	Spacing Policy	200
3.1.4	String Stability	201
3.1.5	Other Objectives	201
3.2	Control Design Methodologies	202
3.2.1	Linear Controller Design	202
3.2.2	Nonlinear Controller Design	203
3.2.3	Model Predictive Control	204
3.2.4	Fuzzy Logic Control	205
4	<i>Conclusion</i>	206

Abstract: The adaptive cruise control (ACC) and cooperative adaptive cruise control (CACC) system is the extension to the conventional cruise control (CC). This chapter focuses on the introduction of various design methodologies for ACC/CACC controllers. The ACC/CACC operation-mode transition and system architecture are presented in detail together with different system components to illustrate concepts and functions of ACC and CACC. A unified control problem formulation and multiple design objectives are then described. Different control design methodologies such as linear control design, nonlinear control design, model predictive control design, and fuzzy control design are reviewed.

1 Introduction

The adaptive cruise control (ACC) system is an enhancement to the conventional cruise control (CC) (ISO15622 2010). Compared with conventional cruise control (CC) systems, which regulate vehicle speed only, an ACC system (Naus et al. 2008) allows drivers to maintain a desired cruise speed as well as a desired following gap with respect to a preceding vehicle if there is no immediate preceding vehicle. The ACC system senses the range (i.e., relative distance) and range rate (i.e., relative speed) to the preceding vehicle with a range sensor (i.e., radar or LIDAR). Such information is used to generate appropriate throttle or brake command to maintain a preset following gap to the preceding vehicle. ACC systems are now commercially available on high-end vehicles of most car manufacturers. They are introduced as one of the advanced driver-assist features that assist the driver's longitudinal control task with limited acceleration range. With reduced workload and stress during daily driving when the ACC system is turned on, the driver could focus more on other important driving tasks and thus achieve improved comfort and safety. From the viewpoint of traffic network operation, the better string stability and tighter following gap of the ACC system compared with manual driving may also provide improvement on traffic safety and capacity with enough penetration rate of ACC systems (Marsden et al. 2001; Vahidi and Eskandarian 2003; Xiao and Gao 2010).

All production-level ACC systems are autonomous in the sense that they can only obtain information about their distance and closing rate to the immediate preceding vehicles using their forward ranging sensors. These sensors are subject to noise, interference, and inaccuracies, which require that their outputs be filtered heavily before being used for control. That introduces response delays and limits the ability of the ACC vehicle to follow other vehicles accurately and respond quickly to changes of traffic flow. For example, experimental results show that the relative speed from the range sensor has about 0.5 s delay compared with the vehicle speed from wireless communication (Bu et al. 2010) which is very significant from the control point of view.

Augmenting the forward ranging sensor data with additional information communicated over a wireless data link from the surrounding vehicles (i.e., vehicle location, speed, acceleration, etc.) and infrastructures (i.e., traffic light states and traffic conditions) makes it possible to overcome these limitations. As an extension to ACC systems, a cooperative

adaptive cruise control (CACC) system further incorporates vehicle-to-vehicle and vehicle-to-infrastructure communication, such as recently developed dedicated short range communication (DSRC), to make use of rich preview information about the surrounding vehicles and environments. Such CACC vehicles can be designed to follow the preceding vehicles with significantly higher accuracy and faster response to changes. Previous research has shown that CACC systems could achieve tighter following gaps and more smooth and “natural” ride in comparison to ACC systems (Arem et al. 2006).

ACC/CACC systems have been studied extensively from highway speed to stop-and-go (ISO22179 2009; Naranjo et al. 2006). Reviews can be found in recent papers (Vahidi and Eskandarian 2003; Xiao and Gao 2010). In general, these research literatures can be classified into the following areas. From the human factor point of view, driver behaviors with ACC/CACC systems were analyzed (Weinberger et al. 2001). The topics of interest include:

- When will the driver turn on/off the ACC/CACC system?
- When will the driver take over control from ACC/CACC system?
- How does the driver’s workload change when the ACC/CACC system is turned on compared with manual driving?
- What is the driver’s choice of following gap in different traffic situations (Novakowski et al. 2010; Viti et al. 2008)?

From the traffic network operation point of view, the influence of ACC/CACC systems on traffic flow characteristics such as safety, capacity, and stability was studied with different penetration rates (Arem et al. 2006). From ACC/CACC system design point of view, the studies of ACC/CACC system development were focused on the development of its two critical components, that is, range sensor and its signal processing algorithm (Shirakawa 2008) and the ACC/CACC controller design.

Since this book is mainly focused on the design of intelligent vehicle and the range sensor technology has already been covered in detail by other chapters of this book, this chapter will primarily focus on the design methodologies of the ACC/CACC controller. The rest of this chapter will be organized as follows: A generic system architecture and operation-mode transition for ACC/CACC systems will be introduced in 🔹 Sect. 2, as well as different system components; the ACC/CACC controller design including control problem formulation, control system structure, control design objectives, and control design methodologies will be presented in 🔹 Sect. 3; the conclusion will be drawn in 🔹 Sect. 4.

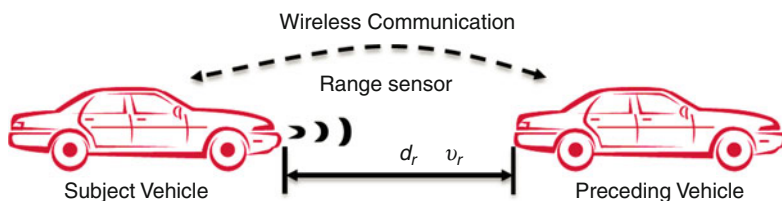
2 System Architecture and Operation Modes

To further understand the design objectives and constraints of the ACC/CACC controller, it is necessary to get a clear picture of the ACC/CACC operation-mode transition and put the ACC/CACC controller in a system perspective. In this section, the operation-mode transition of ACC/CACC system will be introduced first in 🔹 Sect. 2.1. Then a generic system architecture that could achieve those operation goals as well as different system components will be presented in 🔹 Sect. 2.2.

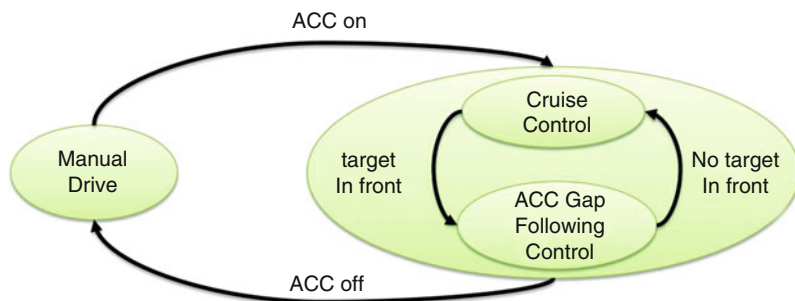
2.1 Overview and Operation Modes

► *Figure 9.1* shows the normal operation of an ACC/CACC system. A vehicle equipped with an ACC or CACC system (subject vehicle) is following its preceding vehicle. The transition of ACC system operation modes is shown in ► *Fig. 9.2*. When the system is turned on, it will regulate the vehicle speed when no preceding vehicle is present, as any conventional cruise control does. Once a preceding vehicle is detected by the range sensor, the system will adjust the vehicle's speed to maintain the gap set by the driver according to the range sensor measurements such as the relative distance d_r and relative velocity v_r , with no control needed from the driver. The driver can take over longitudinal control by either turning off the ACC system or using brake/throttle pedal to override the ACC system commands.

For a CACC system, a wireless communication link needs to be established between at least two vehicles for the data exchange, as shown in ► *Fig. 9.1*. The transition of CACC system operation modes is shown in ► *Fig. 9.3*. When the vehicle in front of the following vehicle is not a CACC preceding vehicle (i.e., the vehicle with wireless communication), ACC mode will be activated. Whenever the CACC preceding vehicle is identified as the vehicle directly in front, the controller enters the CACC mode. The function of “Target ID” mode is to map the position of each vehicle which is directly communicating with the CACC subject vehicle and determine, for example, which communication data stream is



■ **Fig. 9.1**
ACC/CACC vehicle in operation



■ **Fig. 9.2**
ACC system operation-mode transition state machine

coming from the immediate preceding vehicle detected by the range sensor. Bu et al. (2010) implemented a simple pattern recognition algorithm for their small two-car CACC platoon. However, the algorithm may not be suitable for multiple cars in multiple lanes. A more generic solution will require a positioning system that could reliably pinpoint vehicle positions down to the lane level.

2.2 System Architecture and Components

A generic system architecture that is necessary to achieve the aforementioned operational goals is shown in Fig. 9.4. Physically, an ACC/CACC system is usually designed as a distributed system comprised of several different electric control units (ECUs) which are connected by an in-vehicle network such as controller area network (CAN). An ACC/CACC system is composed of the following components or functional blocks:

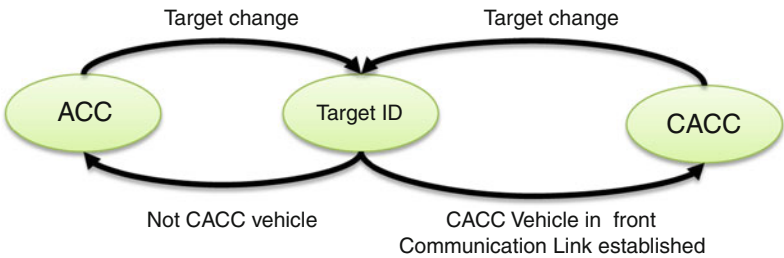


Fig. 9.3
CACC system operation-mode transition state machine

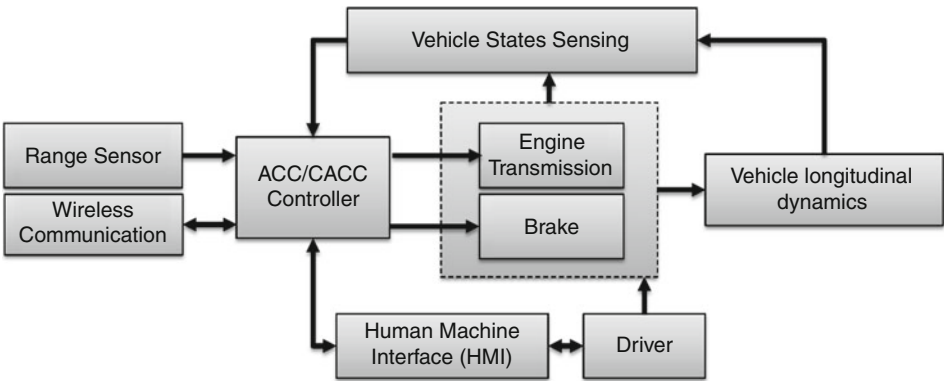


Fig. 9.4
System architecture of an ACC/CACC system

2.2.1 ACC/CACC Controller

The ACC/CACC controller is the center of an ACC/CACC system. It receives driver's commands such as turning on or off system and following gap setting from human-machine interface (HMI). Based on the detection results and measurements from range sensor, internal states measurements from vehicle states sensing (i.e., vehicle speed, gear position, engine speed, brake pressure, etc.), and information about preceding vehicle exchanged through wireless communication, ACC/CACC controller will calculate the corresponding throttle or brake command to maintain the desired gap setting between the ACC/CACC subject vehicle and its preceding vehicle.

2.2.2 Range Sensor

Range sensor is another critical component of the ACC/CACC system. It detects vehicles around the ACC/CACC subject vehicle and measures the relative distance and speed between the subject vehicle and immediate preceding vehicle. The commonly used range sensors are radar, vision sensors, and light detection and ranging sensors (LIDAR). The challenge for the range sensor and its signal processing design is to perform reliable and accurate target detection and measurement under different traffic situations and environmental conditions (Kim and Hong 2004). Although most ACC/CACC systems use forward-looking range sensors, GPS and wireless communication are also used to detect the immediate front vehicle and to measure relative distance and speed (Bruin et al. 2004; Naranjo et al. 2003, 2006).

2.2.3 Vehicle States Sensing

Vehicle states sensing provides vehicle internal states measurements such as wheel speed, engine speed, gear position, and brake pressure. Most of these measurements already exist and are shared through the in-vehicle network of modern vehicles.

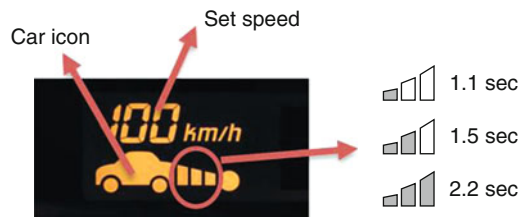
2.2.4 Human-Machine Interface

The driver interacts with the ACC/CACC system through a human-machine interface (HMI). A typical HMI for ACC/CACC system includes displays, switches, and warning devices such as a buzzer. Figure 9.5 shows an exemplar HMI display of a commercial ACC system. When the system is turned on and a preceding vehicle is detected by the range sensor, the car icon which represents the preceding vehicle will be lit and the system enters ACC following gap mode. Three bars behind the preceding car icon represent the following gap setting. In this case, time headway is used for the following gap setting between the subject vehicle and the preceding vehicle. Three bars represent gaps of 1.1s,

1.5s, and 2.2s respectively. When there is no preceding vehicle, the ACC/CACC subject vehicle will cruise at the displayed set speed.

The ON/OFF of ACC system is controlled by switches typically on the steering wheel which are very similar to the switches of conventional cruise control system as shown in [Fig. 9.6](#). One additional cyclic switch is added for the selection of the following gap setting.

Other HMI features include different buzzer sounds to warn the driver that the following gap is too close and driver needs to take over longitudinal control, or to remind the driver that the ACC system is shut off automatically due to low speed. As suggested by ISO standards (ISO15622 [2010](#); ISO22179 [2009](#)), brake light needs to be lit if brake is applied by the ACC/CACC system.



■ Fig. 9.5
Display of a commercial ACC system



■ Fig. 9.6
Control switches of a commercial ACC system on steering wheel

2.2.5 Wireless Communication

To exchange information among the subject vehicle and other vehicles for the CACC operation, wireless communication links need to be established among vehicles. Preceding vehicles' and surrounding vehicles' internal states such as speed, acceleration, gear position, and accelerator or brake pedal actuation can be sent to the following subject vehicle to enable CACC control. Wireless communication link could also be established between infrastructure and the CACC subject vehicle. Information such as traffic light and future traffic condition can be sent to the subject vehicle to further optimize CACC controller performance.

3 ACC/CACC Controller Design

The ACC/CACC controller is a critical component of the ACC/CACC system. The focus of this section is to review different design methodologies of the ACC/CACC controller. To facilitate the illustration, control problem formulation and design objectives for the ACC/CACC controller design will be presented first in [Sect. 3.1](#). Different control design methodologies will be reviewed in the following sections.

3.1 Control Problem Formulation and Design Objectives

The purpose of this section is to provide a unified control problem formulation and to present controller design objectives for the illustration of different design methodologies in the existing literatures.

3.1.1 Control Problem Formulation

A vehicle platoon equipped with either an ACC or CACC system is shown in [Fig. 9.7](#). x_i , \dot{x}_i , and \ddot{x}_i represent front bumper position, speed, and acceleration of i th vehicle. For vehicle i , its preceding vehicle is vehicle $i - 1$ and its following vehicle is vehicle $i + 1$. The relative distance $x_{r,i}$ and velocity $\dot{x}_{r,i}$ are defined as

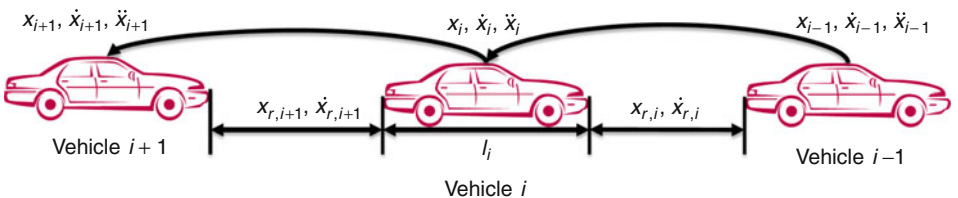


Fig. 9.7
ACC/CACC vehicle platoon

$$\begin{cases} x_{r,i}(t) = x_i(t) - x_{i-1}(t) - l_{i-1} \\ \dot{x}_{r,i}(t) = \dot{x}_i(t) - \dot{x}_{i-1}(t) \end{cases} \quad (9.1)$$

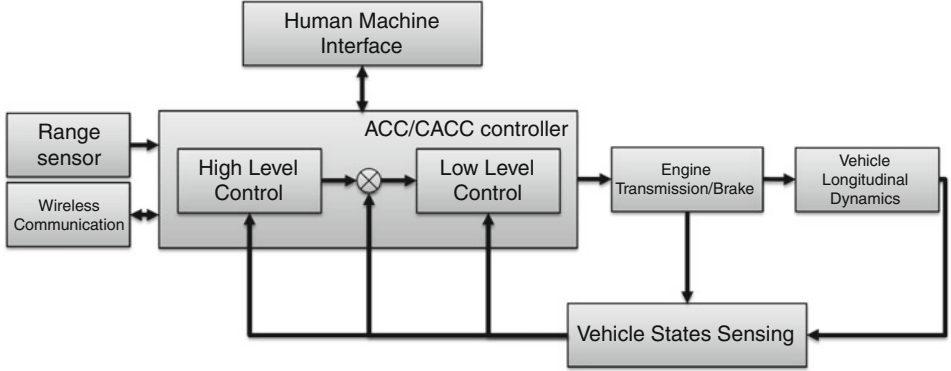
where l_{i-1} represents the length of vehicle $i-1$. Given the desired gap setting $x_{rd,i}$ for the i th vehicle, one of the control objectives is to synthesize appropriate control command u (either throttle or brake) so that the gap regulating error or spacing error e_i is minimized.

$$e_i(t) = x_{r,i}(t) - x_{rd,i}(t) \quad (9.2)$$

CACC vehicles could communicate with each other if vehicles are in the range of wireless communication. Using information from multiple preceding vehicles could provide better performance for CACC system. For example, string stability could be achieved for constant spacing policy if information from platoon leader and nearest preceding vehicle is available to every CACC vehicle in the platoon (Swaroop and Hedrick 1996). Other communication configurations include using information from both immediate front and back vehicles (Zhang et al. 1999). However, using information from multiple vehicles in the platoon for the CACC controller design will make the system more complex and less feasible for implementation in real traffic environment. Figure 9.7 shows a “semi-autonomous” (Rajamani and Zhu 2002) or a “decentralized” (Naus et al. 2010b) system configuration in which each vehicle only receives information from the immediate front vehicle. Such a system configuration is easy to implement and still preserves most performance advantages as shown in the latter review.

3.1.2 Control System Structure

As shown in Fig. 9.8, a hierarchical control system structure which consists of an inner loop (“servo loop”) and an outer loop controller is usually adopted for the vehicle’s longitudinal control purpose including ACC/CACC control (Li et al. 2011). The outer loop controller, also called the high-level controller, synthesizes a desired speed v_d (Bu et al. 2010; Naranjo et al. 2003) or acceleration command a_d according to the driver gap setting and the sensing information from range sensor, wireless communication, and internal vehicle states. If a vehicle follows such speed or acceleration commands closely, the control objectives such as gap error regulation can be achieved. Therefore, the inner loop controller or the low-level controller will generate corresponding throttle or brake commands so that the vehicle will follow the desired speed or acceleration profile generated by the high-level controller. Both inner speed loop and inner acceleration loop designs can be found in commercial ACC systems and academic literatures. Low-level speed controller can use high-resolution wheel speed signals as feedback measurement and does not require a very accurate vehicle longitudinal model. Since either direct measurement of vehicle longitudinal acceleration is lacking or direct measurement is usually contaminated by local vibration noise, low-level acceleration controller typically requires an accurate vehicle longitudinal model including engine, transmission, and brake.



■ Fig. 9.8
Hierarchical structure of ACC/CACC control system

To simplify the design of the high-level control, the closed low-level loop is usually approximated by a first-order system:

$$\begin{cases} \ddot{x} = \frac{K_a}{\tau_a s + 1} a_d \\ \dot{x} = \frac{K_v}{\tau_v s + 1} v_d \end{cases} \quad (9.3)$$

where K_a , τ_a , K_v , and τ_v represent static gains and time constants for low-level acceleration control and low-level speed control, respectively.

3.1.3 Spacing Policy

How to determine the desired gap, x_{rd} , is the first question usually asked when an ACC/CACC system is designed. Constant spacing policy (Shladover 1978) where $x_{rd} = d_0$ and d_0 is a constant was proposed for the vehicle platoon of Automatic Highway System (AHS). It is shown that string stability can be achieved only if the lead vehicle's velocity or velocity and acceleration are broadcasted to the other vehicles in the platoon. A speed-dependent spacing policy also called time headway is the most commonly used for commercial ACC systems. The desired spacing can be defined as

$$x_{rd} = d_0 + \dot{x} t_{hw} \quad (9.4)$$

where d_0 is a constant and represents the minimum safe distance and t_{hw} is the time headway or time gap. Time headway policy is more similar to a driver's daily experience. Furthermore, a simple controller (Rajamani and Zhu 2002) that only requires speed measurements of the immediate front vehicle could guarantee string stability. One of the drawbacks of the time headway policy is its poor robustness against traffic flow fluctuation (Junmin and Rajamani 2004; Zhou and Peng 2005). Spacing policies that

could preserve traffic flow stability in all traffic situations were designed. Junmin and Rajamani (2004) developed a nonlinear spacing policy for the stability of traffic flow:

$$x_{rd,i} = \frac{1}{\rho_m \left(1 - \frac{\dot{x}_i}{v_f}\right)} \quad (9.5)$$

where ρ_m is a traffic density parameter and v_f is a vehicle speed parameter. Zhou and Peng (2005) designed a spacing policy that included a quadratic term of vehicle speed and was optimized for both string stability and traffic flow stability:

$$x_{rd,i} = 3 + 0.0019\dot{x}_i + 0.0448\dot{x}_i^2 \quad (9.6)$$

Other spacing policies were also proposed for different scenarios. A complex nonlinear spacing policy was designed for a vehicle platoon comprised of commercial heavy-duty vehicles (Yanakiev and Kanellakopoulos 1998). A safe distance policy was presented to include braking capability of the following vehicle (Zhang et al. 1999):

$$x_{rd,i} = h_1(\dot{x}_i^2 - \dot{x}_{i-1}^2) + h_2\dot{x}_i + h_3 \quad (9.7)$$

where h_1 , h_2 , and h_3 are positive constants that depend on the following vehicle's braking capability.

3.1.4 String Stability

The rigorous mathematical definition of string stability of vehicle platoon can be found in earlier work (Swaroop and Hedrick 1996). Practically, it means that gap regulation errors will not be amplified from the lead vehicle to the last vehicle in the platoon. A necessary condition in the frequency domain can be expressed as follows (Naus et al. 2010b; Rajamani and Zhu 2002). Define the relationship of the gap regulation errors of adjacent vehicles in the platoon as

$$H(s) = \frac{e_i(s)}{e_{i-1}(s)} \quad (9.8)$$

where $e_i(s)$ is the Laplace transform of vehicle i 's gap regulation error. Then, the string stability of vehicle platoon will be guaranteed if the following condition is satisfied:

$$\|H(s)\|_\infty < 1 \quad (9.9)$$

From the traffic network operation point of view, string stability means smooth traffic flow and less “shock wave.” From the driver's point of view, guaranteed string stability will provide smooth ride and possible safety benefit (Lu et al. 2002).

3.1.5 Other Objectives

Other design objectives may include ride comfort, fuel consumption (Bageshwar et al. 2004; Corona and De Schutter 2008; Corona et al. 2006; Li et al. 2011), and even stop time before

traffic light (Asadi and Vahidi 2011). Ride comfort is usually quantified by adding limits on longitudinal acceleration and its derivative, longitudinal jerk. Fuel consumption can be derived from vehicle longitudinal model. To simplify the fuel consumption model, Li et al. (2011) assume that smooth acceleration will usually result in low fuel consumption.

Driver acceptance is also an important design objective for the ACC/CACC system. The assumption is that the dynamic characteristics of the ACC/CACC system should not be very different from that of a driver's longitudinal driving. An average longitudinal driving model is typically added in the ACC/CACC controller to generate the desired dynamic characteristics (Bageshwar et al. 2004; Li et al. 2011; Moon and Yi 2008; Naus et al. 2010a; Persson et al. 1999).

3.2 Control Design Methodologies

Different control methodologies have been proposed in research literatures for the design of the ACC/CACC controller. Linear control design will be used as a starting point for illustrations. Linear controller alone could not achieve adequate performance for all traffic conditions. To design an ACC/CACC controller that has satisfactory performance in real traffic environment, either pure linear controller needs to be complemented by certain nonlinear elements or a nonlinear design needs to be adopted. The ACC/CACC controller design often needs to meet multiple contradictory design objectives with stringent constraints. Model predictive control (MPC) is a control framework that could perform constrained multi-objective optimization. Finally, fuzzy logic controller is introduced to mimic human driver behavior for the ACC controller design.

It is worth noting that most design methodologies reviewed here are from academic literatures. Although ACC systems have been commercially available for a long time, the approaches of ACC controllers taken by car manufacturers or OEM manufacturers are still not publicly revealed.

3.2.1 Linear Controller Design

As mentioned in Sect. 3.1.2, the low-level closed loop can be approximated by a first-order linear system with a well-designed low-level controller when a hierarchical controller structure is adopted. Various linear control design techniques such as PID, linear optimal control, and gain scheduling can then be applied to synthesize the high-level controller.

A very simple linear controller was proposed (Rajamani and Zhu 2002) with guaranteed string stability for the ACC controller design with time headway policy. The synthesized desired acceleration can be expressed as

$$a_{d,i} = -\frac{1}{h(\dot{x}_{r,i} + \lambda e_i)} \quad (9.10)$$

PID controllers with fixed gain and adaptive gain were studied (Ioannou et al. 1993) and produced satisfactory results for both high-fidelity model simulation and

experimental study with time headway ranging policy for the ACC system. Persson et al. (1999) designed a PI controller augmented with an average linear driver model to achieve normal driving characteristics for ACC system. Standard PD controller (Naus et al. 2010b) was designed for the ACC controller and proved to be string stable for the constant time headway policy. Based on the PD control design for the ACC controller, the CACC controller is designed by filtering preceding vehicle's acceleration as a feed-forward term. The designed CACC controller is also proven to be string stable.

Liang and Peng (1999) proposed a linear optimal design for ACC high-level controller with a constant time headway policy. Quadratic functions of weighted range and range rate error were used as the performance index. String stability is guaranteed with such control problem formulation. Using the same design method, Moon et al. (2009; Moon and Yi 2008) tuned a weighting matrix of performance index to achieve naturalistic driving behavior. Manual driving data shows that drivers tend to use large acceleration in low speed while driving smoothly in high speed. Therefore, variable acceleration limits at different driving speeds were applied to the synthesized desired acceleration a_d .

Wildmann et al. (2000) designed a gain-scheduled linear controller to enable ACC function engaged between 30 and 160 km/h. As the authors pointed out, in general, a pure linear control design could provide acceptable performance for steady-state vehicle following. However, in other situations where traffic pattern changes abruptly, that is, a vehicle cut-in, preceding vehicle lane change, or preceding vehicle braking/accelerating hard, performance is not adequate. This is due to the limitation of linear control design. Wildmann et al. used additional nonlinear elements to compensate the drawback of pure linear design. For example, nonlinear gain was added to force a quicker brake action when vehicle gap is shortened abruptly by the hard braking of the preceding vehicle or cut-in of other vehicles. When the following gap is suddenly enlarged due to lane change of subject vehicle or preceding vehicle, certain open loop acceleration profile is used for “catching up” before the engagement of high gain gap regulating controllers to avoid controller windup and jerky motions.

3.2.2 Nonlinear Controller Design

Lu et al. (2002) applied sliding mode control to the design of ACC and CACC controller. Two kinds of slide surfaces S_1 and S_2 were proposed for a constant time headway policy

$$\begin{aligned} S_1 &= (\dot{x}_r - \bar{x}t_{hw}) + k_1 e \\ S_2 &= \dot{x}_r + k_2 e \end{aligned} \quad (9.11)$$

where k_1 and k_2 are positive parameters. Both sliding surface designs will lead to a globally stable closed loop system. The design can be applied to both ACC and CACC controller designs.

Martinez and Canudas-de-Wit (2007) proposed a reference model-based nonlinear control approach for ACC controller design. A nonlinear reference model adopted was designed to take into account both driver's characteristics and safety requirements. The reference model acts as a feed-forward term to the ACC feedback controller.

3.2.3 Model Predictive Control

As mentioned in Sects. 3.1.3–3.1.5, there are multiple design objectives in the ACC/CACC controller design such as minimizing gap regulating error, preserving string stability, increasing driver riding comfort, and minimizing fuel consumption. Some of these objectives are often contradictory. For example, frequent acceleration and deceleration will degrade driver ride comfort but is necessary to maintain an accurate gap distance between the preceding vehicle and the following vehicle especially when preceding vehicle's motion is not smooth. The design of the ACC/CACC controller also has lots of hard constraints such as actuator (brake and throttle) limit and safety limit. In the previously reviewed design methods, trying to meet all the design objectives within design constraints is usually done by tuning a controller's parameters in a trial-and-error way. Model predictive control (MPC) is a control framework that can optimize multiple performance criteria under different design constraints. The formulation of MPC usually results in a constrained optimization problem, which can be solved by multiple solvers. Due to its complexity of computation, MPC used to be applied for chemical process control where plant dynamics is slow and real-time computation requirement is not that stringent. With the development of fast computing devices, MPC is extending its application to the areas like automotive control where fast response is necessary. In recent years, there are growing interests in applying MPC design methodology in the ACC/CACC controller design. Recent literatures show that the MPC method has been successfully implemented in different aspects of the ACC/CACC controller design, and it is reviewed below.

ACC control is composed of two phase operations: transitional and steady state. When a new preceding vehicle is identified with large gap error, the subject vehicle with the ACC system must either slow down or speed up immediately to enlarge or reduce gap distance to the vicinity of gap setting. This is called the transitional process. Bageshwar et al. (2004) designed a controller using MPC for the transitional maneuvers of ACC vehicles. The advantage of the MPC framework is that the same problem formulation can be used to calculate control law for all feasible initial conditions. The control objectives and constraints such as beginning steady-state control phase with zero-range rate, collision avoidance, and acceleration limit can be incorporated into the design problem explicitly.

Naus et al. (2010a) presented a systematic way for the design and tuning of ACC controller. The key characteristics of a designed ACC system are several parameters which are convenient and intuitive to tune. To incorporate multiple design objectives and constraints, an MPC framework was adopted. Since online optimization of MPC was usually time consuming and might not be suitable for real-time implementation, an

explicit MPC approach was used to solve the optimization problem offline and generate an analytical solution.

Corona and Shutter (2008) did not use the hierarchical control structure introduced in Sect. 3.1.2. They synthesized a single ACC controller which outputs brake, throttle, and gear shifting commands with the MPC design method. Since a vehicle longitudinal model including gear transmission is a hybrid model, mixed integer programming is used to solve the optimization problem.

Li et al. (2011) presented rather complete design objectives and constraints and set up an MPC framework for the solution. The objective list included gap regulation error, fuel economy, ride comfort, and driver desired characteristics. To simplify the resulting optimization problem, restrictions on the acceleration level instead of fuel consumption were formulated into the control problem with the assumption that fuel consumption is mainly dominated by vehicle acceleration levels. An average driver model was incorporated into the design problem to generate driver desired dynamic characteristics. Limits on vehicle longitudinal acceleration and jerk were applied to enforce driver ride comfort requirements. A weighted quadratic function of regulate gap error and relative speed error was used as performance index for gap regulating error.

Asadi and Vahidi (2011) described an interesting application of wireless communication for CACC systems. A CACC vehicle could communicate with the infrastructure to get information such as current and future traffic signal states. That information can be used in an MPC framework to minimize waiting time before traffic signal and improve fuel economy.

Bu et al. (2010) reported the first operation of a CACC system on the public road with extensive mileages. Two Infinity FX45s with factory-installed ACC systems were retrofitted with the CACC prototype system designed by authors. A low-level speed servo loop was designed to compensate the unknown vehicle longitudinal model with frequency shaping and gain scheduling methods. The high-level control was formulated into an MPC problem to satisfy stringent performance requirements and multiple design constraints. An indirect adaptive method was used to identify parameters related to the preceding vehicle. Extensive testing on vehicle proving ground and public highway showed that the designed CACC system could maintain a 0.6 s time gap consistently and reliably. The testing also showed that the designed CACC system could provide better performance compared with factory install ACC system in term of string stability

3.2.4 Fuzzy Logic Control

Naranjo et al. (2003, 2006) proposed an adaptive fuzzy logic controller for the ACC controller design. In order to emulate human driving, the design process started from generalizing fuzzy reasoning rules for human driving. Fuzzy PID controllers were designed for both high-level control and low-level acceleration servo control. The results could be applied to both highway speed driving and urban driving with frequent stop-and-go's.

4 Conclusion

Adaptive cruise control (ACC) and cooperative adaptive cruise control (CACC) systems are extensions to the conventional cruise control (CC). This chapter focuses on the introduction of various design methodologies for the ACC/CACC controller. To help the illustrations, ACC/CACC operation-mode transition and system architecture are presented in detail together with different system components. A unified control problem formulation as well as a hierarchical control system structure and multiple design objectives are introduced. Different control design methodologies such as linear control design, nonlinear control design, model predictive control design, and fuzzy logic control design are reviewed extensively.

Although ACC has been commercially available for years, the full deployment of commercial CACC system is still years away. Aside from the design and testing of CACC controller, there are still road blocks such as a standardized communication protocol for CACC system, a reliable lane-level vehicle positioning system, and maps for vehicle “identification” purpose. These remain topics of further research and development.

References

- Asadi B, Vahidi A (2011) Predictive cruise control: utilizing upcoming traffic signal information for improving fuel economy and reducing trip time. *IEEE Trans Control Syst Technol* 19: 707–714
- Bageshwar VL, Garrard WL, Rajamani R (2004) Model predictive control of transitional maneuvers for adaptive cruise control vehicles. *IEEE Trans Veh Technol* 53:1573–1585
- Bruin Dd, Kroon J, Klaveren Rv, Nelisse M (2004) Design and test of a cooperative adaptive cruise control system. In: *Proceedings of 2004 IEEE intelligent vehicle symposium*, pp 392–396. Parma
- Bu F, Tan H-S, Jihua H (2010) Design and field testing of a cooperative adaptive cruise control system. In: *Proceedings of the 2010 American control conference*, pp 4616–4621. Baltimore
- Corona D, De Schutter B (2008) Adaptive cruise control for a SMART car: a comparison benchmark for MPC-PWA control methods. *IEEE Trans Control Syst Technol* 16:365–372
- Corona D, Lazar M, De Schutter B, Heemels M (2006) A hybrid MPC approach to the design of a Smart adaptive cruise controller. In: *Proceedings of the 2006 IEEE international conference on control applications*. Munich, pp 231–236
- Ioannou P, Xu Z, Eckert S, Clemons D, Sieja T (1993) Intelligent cruise control: theory and experiment. In: *Proceedings of the 32nd IEEE conference on decision and control*, vol 2. San Antonio, pp 1885–1890
- ISO15622 (2010) Intelligent transport systems – Adaptive Cruise Control systems – Performance requirements and test procedures
- ISO22179 (2009) Intelligent transport systems – Full speed range adaptive cruise control (FSRA) systems – Performance requirements and test procedures
- Junmin W, Rajamani R (2004) Should adaptive cruise-control systems be designed to maintain a constant time gap between vehicles? *IEEE Trans Veh Technol* 53:1480–1490
- Kim Y-S, Hong K-S (2004) An IMM algorithm for tracking maneuvering vehicles in an adaptive cruise control environment. *Int J Control Autom Syst* 2:310–318
- Li S, Li K, Rajamani R, Wang J (2011) Model predictive multi-objective vehicular adaptive cruise control. *IEEE Trans Control Syst Technol* 19:556–566
- Liang C-Y, Peng H (1999) Optimal adaptive cruise control with guaranteed string stability. *Veh Syst Dyn: Int J Veh Mech Mobil* 32:313–330

- Lu X-Y, Hedrick JK, Drew M (2002) ACC/CACC – control design, stability and robust performance. In: American control conference. Anchorage
- Marsden G, McDonald M, Brackstone M (2001) Towards an understanding of adaptive cruise control. *Transp Res Part C: Emerg Technol* 9:33–51
- Martinez J-J, Canudas-de-Wit C (2007) A safe longitudinal control for adaptive cruise control and stop-and-go scenarios. *IEEE Trans Control Syst Technol* 15:246–258
- Moon S, Yi K (2008) Human driving data-based design of a vehicle adaptive cruise control algorithm. *Veh Syst Dyn: Int J Veh Mech Mobil* 46:661–690
- Moon S, Moon I, Yi K (2009) Design, tuning, and evaluation of a full-range adaptive cruise control system with collision avoidance. *Control Eng Pract* 17:442–455
- Naranjo JE, Gonzalez C, Reviejo J, Garcia R, de Pedro T (2003) Adaptive fuzzy control for inter-vehicle gap keeping. *IEEE Trans Intel Transp Syst* 4: 132–142
- Naranjo JE, Gonzalez C, Garcia R, de Pedro T (2006) ACC + Stop&go maneuvers with throttle and brake fuzzy control. *IEEE Trans Intel Transp Syst* 7:213–225
- Naus G, Ploeg J, van de Molengraft R, Steinbuch M (2008) Explicit MPC design and performance-based tuning of an adaptive cruise control stop-&-go. In: *IEEE intelligent vehicles symposium*. Eindhoven, pp 434–439
- Naus GJL, Ploeg J, Van de Molengraft MJG, Heemels WPMH, Steinbuch M (2010a) Design and implementation of parameterized adaptive cruise control: An explicit model predictive control approach. *Control Eng Pract* 18:882–892
- Naus GJL, Vugts RPA, Ploeg J, van de Molengraft MJG, Steinbuch M (2010b) String-stable CACC design and experimental validation: a frequency-domain approach. *IEEE Trans Veh Technol* 59:4268–4279
- Novakowski C, Shladover S, Bu F, O’Connell J, Spring J, Dickey S, Nelson D. (2010) Cooperative adaptive cruise control: testing drivers’ choices of following distances. California PATH, Richmond
- Persson M, Botling F, Hesslow E, Johansson R (1999) Stop and go controller for adaptive cruise control. In: *Proceedings of the 1999 IEEE international conference on control applications*. Kohala Coast, Vol 2, pp 1692–1697, Vol 1692
- Rajamani R, Zhu C (2002) Semi-autonomous adaptive cruise control systems. *IEEE Trans Veh Technol* 51:1186–1192
- Shirakawa K (2008) PRISM: an in-vehicle CPU-oriented novel azimuth estimation technique for electronic-scan 76-GHz adaptive-cruise-control radar system. *IEEE Trans Intel Transp Syst* 9:451–462
- Shladover S (1978) Longitudinal control of automated guideway transit vehicles within platoons. *ASME J Dyn Syst, Meas Control* 100:302–310
- Swaroop D, Hedrick JK (1996) String stability of interconnected systems. *IEEE Trans Autom Control* 41:349–357
- Vahidi A, Eskandarian A (2003) Research advances in intelligent collision avoidance and adaptive cruise control. *IEEE Trans Intel Transp Syst* 4:143–153
- Van Arem B, Van Driel JG, Visser R (2006) The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans Intel Transp Syst* 7:429–436
- Viti F, Hoogendoorn SP, Alkim TP, Bootsma G (2008) Driving behavior interaction with ACC: results from a field operational test in the Netherlands. In: *Intelligent vehicles symposium, 2008 IEEE*. Eindhoven, pp 745–750
- Weinberger M, Winner H, Bubbl H (2001) Adaptive cruise control field operational test—the learning phase. *JSAE Review* 22:487–494
- Widmann GR, Daniels MK, Hamilton L, Humm L, Riley B, Schiffmann JK, Schnelker DE, Wishon W (2000) Comparison of lidar-based and radar-based adaptive cruise control systems. In: *SAE 2000 world congress*, Detroit
- Xiao L, Gao F (2010) A comprehensive review of the development of adaptive cruise control systems. *Veh Syst Dyn: Int J Veh Mech Mobil* 48:1167–1192
- Yanakiev D, Kanellakopoulos I (1998) Nonlinear spacing policies for automated heavy-duty vehicles. *IEEE Trans Veh Technol* 47:1365–1377
- Zhang Y, Kosmatopoulos EB, Ioannou PA, Chien CC (1999) Autonomous intelligent cruise control using front and back information for tight vehicle following maneuvers. *IEEE Trans Veh Technol* 48:319–328
- Zhou J, Peng H (2005) Range policy of adaptive cruise control vehicles for improved flow stability and string stability. *IEEE Trans Intel Transp Syst* 6:229–237

10 Vehicle Lateral and Steering Control

Damoon Soudbakhsh · Azim Eskandarian
Center for Intelligent Systems Research, The George Washington
University, Washington, DC, USA

- 1 *Introduction*210**

- 2 *Major Vehicle Components in Lateral Motion*212**
 - 2.1 Tires 212
 - 2.2 Linear Tire Model for Uniform Normal Force Distribution 212
 - 2.3 Steering Wheel 213
 - 2.4 Suspension System 215

- 3 *Vehicle Model*215**
 - 3.1 Tire Models 217
 - 3.1.1 Magic Formula Tire Model 217
 - 3.1.2 Brush Model 218
 - 3.1.3 Swift Tire Model 219
 - 3.1.4 Linear Tire Model 219
 - 3.2 Error Coordinates 220
 - 3.3 Global Position of the Vehicle 221
 - 3.4 Using Yaw Rate and Slip Angle as State Space Variables 221

- 4 *Controllers*223**

- 5 *Desired Trajectory for Lane-Change and Lane Keeping Maneuvers*224**

- 6 *Some Experimental Studies*226**

- 7 *Case Studies*227**

- 8 *Electronic Stability Control (ESC)*228**

- 9 *Conclusions*229**

Abstract: Run-off-road crashes are responsible for about 34% of road fatalities. In addition, side swiping crashes are also caused by either unintended or ill-timed lane departures. Many systems have been developed to reduce these kinds of accidents. Vehicle lateral control and warning systems are introduced for this reason. Vehicle lateral control systems can help the drivers to change their lanes safely or assist them in parking or even performing evasive maneuvers. This chapter aims to serve as an introduction to dynamics, control, and modeling of vehicle lateral motion in the context of intelligent vehicles. First, the major components that determine lateral dynamics of the vehicle are briefly described. Then, the basics of tire dynamics and vehicle modeling are presented. Many different linear and nonlinear controllers have been suggested in the literature to control the combined lateral and longitudinal motion of the vehicle. It is expected that in the future vehicle lateral control will be extended to performing lane change or evasive maneuvers. A brief review of the control methods and some of the experimental implementations of autonomous trajectory following (lane change maneuvers) are presented. A few examples of their commercial productions are provided. Finally, some concluding remarks about the future of vehicle lateral control and the current state of the development in this area are given.

1 Introduction

Data from Fatal Analysis Reporting System (FARS) shows that run-off-road (ROR) crashes were responsible for 34% of the total fatalities in 2008. Aggregating all types of lane departure crashes, roadway departures were responsible for about 53% of all road fatalities (FHWA Retrieved on Apr 7 2010).

Deviation from the lane may occur due to various external factors, vehicle control loss, or driver errors. In case of obstacle avoidance, experienced drivers can perform evasive maneuvers and stabilize their vehicle after the maneuver, whereas drivers with less experience have difficulties in stabilizing their vehicle and may be involved in the lane departure accident following an evasive maneuver (Maeda et al. 1977).

The next generation of major vehicular safety enhancements is expected from active safety systems, which assist the drivers in control functions. Steering assistance can take various forms. Among these are the lane-keeping support systems, which have produced notable research results (Kawazoe et al. 2001). Lane keeping is accomplished by steering and lateral control of the vehicle. These systems mainly differ in the controller used or in the control input signals. The systems typically either use (1) steering angle as control input, which is called steering angle control, and (2) steering torque as the control input, called steering torque control (Kawazoe et al. 2001; Nagai et al. 2002). Most of the systems developed for lane keeping have used generated steering torque to control the vehicle (Montiglio et al. 2006; Ishida et al. 2004; Pohl and Ekmark 2003). However, many researchers have chosen steering angle control over steering torque control (Nagai et al. 2002). In the real world, the actual lane position can be estimated by using sensors such as an in-vehicle video camera and a signal processor to estimate the position of the car relative to the lane markings (Montiglio et al. 2006; Ishida et al. 2004; Pohl and Ekmark 2003).

The major function of vehicle lateral control system are lane following, lane change maneuver, and collision avoidance (Hedrick et al. 1994). Lane change maneuver can be divided into four subfunctions of sensing, perceiving, deciding, and acting. Passive sensors such as cameras, active sensors such as radar, laser, and near infrared cameras, and a combination of these sensors (sensor fusion) can be used to detect lanes and obstacles (Tideman et al. 2007). A variety of algorithms have been used by researchers to assess safety of the maneuver. For the case of lane change warning, many algorithms based on (1) Current position of the driver; (2) Time to cross the line; and (3) interpretation of road scene have been proposed (Tideman et al. 2007).

Many of these systems have been commercialized such as a haptical lane-keeping support system for heavy trucks, aimed to prevent unwanted lane departures. In this system, the controller uses heading angle and lateral position to generate the steering torque demand necessary to maintain the vehicle in its lane (Montiglio et al. 2006). Another example is the combination of adaptive cruise control system (ACC) and lane-keeping assistant system (LKAS), which uses the vehicle' heading angle and the lateral deviation within the lane to generate a steering torque to be combined with the driver's steering torque (Ishida et al. 2004).

Although there are still problems in terms of false alarm and intervention failure in active systems, the future of these systems looks promising (Andreas et al. 2007). There are many new systems that can be implemented in the vehicles, which can make autonomous tasks such as changing lanes, parking, and maneuvering through obstacles possible (Leonard et al. 2009). Some of the active safety systems that can be implemented in the vehicles as of now as listed in Andreas et al. (2007) are: (1) LKA (Lane-Keeping Aid system) which uses cameras to detect lanes and uses a steering wheel actuator to keep the vehicle in its lane, (2) LCA (Lane Change Assist) warns drivers of another vehicle in the blind spot if they start changing lane, (3) Automatic braking when forward collision is unavoidable, (4) Curvature Warning if the approaching curve is too sharp for the current speed of the vehicle, (5) Scanners such as wildlife or pedestrian scanner which warn the driver if these objects enter the road in front of the vehicle.

The rest of this chapter is organized as follows: first the major components of the vehicle in lateral motion are presented. In this section, the dynamics of tire forces are explained using a simplified tire model. ➤ Section 3 describes a vehicle bicycle model which is the most common model in analyzing vehicle lateral motion. In this section, some of the most famous tire models and their features are introduced, and the state space form of the vehicle model using a linear tire model is provided. Then this vehicle model is presented in different coordinates (errors coordinates, global coordinates, and yaw and slip angles coordinate systems). In the following ➤ Sects. 4 and ➤ 5, a review of control methods used for vehicle lateral control as well as some of the best known desired trajectories for lane change maneuvers are presented. In ➤ Sect. 6, some experimental studies in the academic world are presented, and in ➤ Sect. 7, some of the lane-keeping systems that are already available in the vehicles are reviewed. ➤ Section 8 is a short description of electronics stability control (ESC) and its potential benefits as it is related to vehicle lateral control. Finally, ➤ Sect. 8 provides a summary and some concluding remarks of this chapter.

2 Major Vehicle Components in Lateral Motion

Although many parts of the vehicle can contribute to the lateral dynamics of the vehicle; here the most important ones are presented: (1) Tires, (2) Steering system, and (3) suspension system. It should be noted that braking and acceleration can also significantly change the response of the vehicle since they result in a load transfer (from rear to front and vice versa). However, to represent simple dynamics of the vehicle, it is assumed that the forward velocity of the vehicle remains constant through the maneuver and there is no braking or acceleration involved.

2.1 Tires

Tire lateral forces are the major forces applied to the vehicle during the maneuver. These lateral forces are generated when the tire advances in the direction of travel while having different heading direction. In this kind of situation, the elements in the contact patch stay in their original position, and are deflected sideways with respect to the tire. In this process, lateral force is increased as the element moves rearward up to a point where slip occurs (Gillespie 1992). The resultant between the direction of tire heading and the direction of the travel is called slip angle (Gillespie 1992). The maximum value of the lateral friction force is μF_z , where μ is the tire-road friction coefficient, and F_z is the normal force on the tire. When longitudinal accelerations/decelerations are considered, the maximum forces can be estimated using combined slip models and friction circle.

2.2 Linear Tire Model for Uniform Normal Force Distribution

To better understand the mechanism of lateral tire force generation a simple model of the tire known as elastic foundation model (Fig. 10.1) is presented in this section. This tire model assumes that contact patch elements act independently (Rajamani 2006).

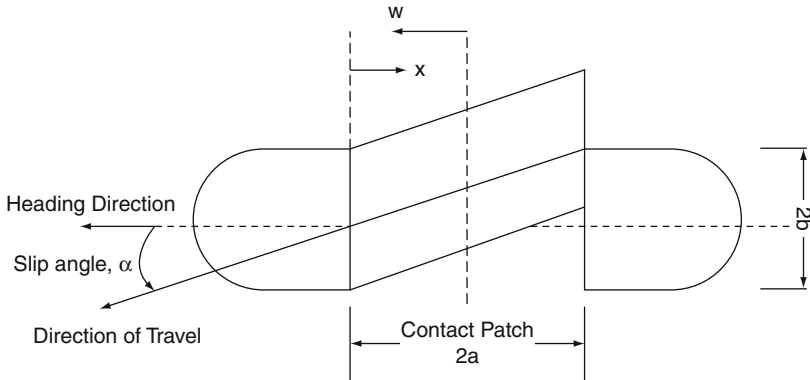
Let c to be the lateral stiffness per unit length of the tire, and $\gamma(x)$ be the lateral displacement of the tire (Fig. 10.2). With the assumptions of uniform normal pressure over contact patch, and small slip angle ($\gamma(x) = \tan(\alpha) \cdot x \approx \alpha \cdot x$), then total lateral force (F_y), and Self-aligning moment are given by the following expressions:

$$dF = c \cdot \gamma(x) \cdot dx \xrightarrow{\text{yields}} F_y = \int_0^{2a} c \cdot \gamma(x) dx = 2ca^2\alpha = C_\alpha \alpha \quad (10.1)$$

$$M_z = \int_0^{2a} c \cdot \gamma(x) \cdot (x - a) dx = \frac{F_y(2a)}{6} = \frac{a}{3} F_y \quad (10.2)$$

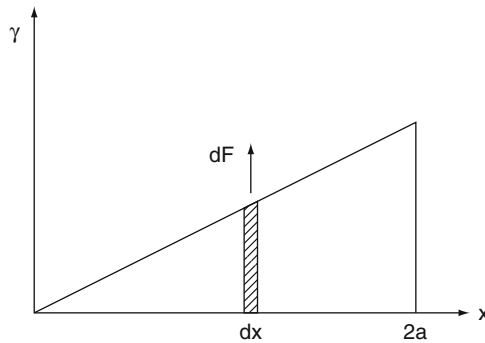
C_α is called cornering stiffness.

A broader selection of linear and nonlinear tire models is given in Sect. 3.1.



■ Fig. 10.1

Rolling tire deformation at small slip angles (Reproduced by permission from (Rajamani 2006))



■ Fig. 10.2

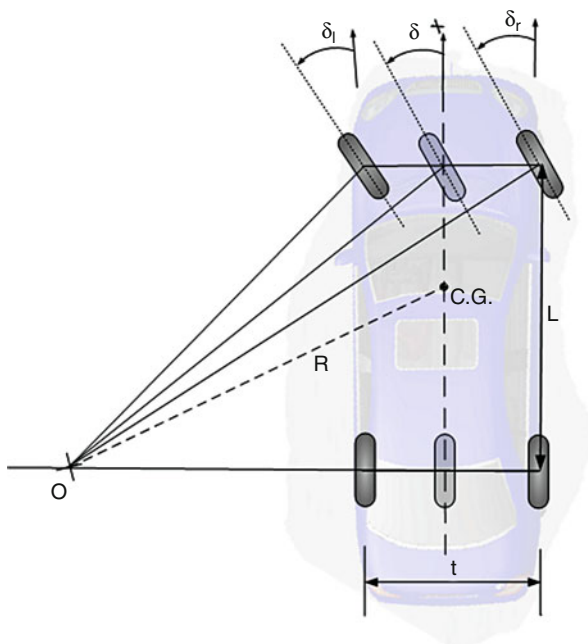
Tire deformation notation

2.3 Steering Wheel

There are several steering system designs; however, they are functionally similar (Gillespie 1992). Two of the most common systems for passenger cars are: Rack-and-Pinion Linkage and Steering gearbox.

► Figure 10.3 shows the geometry of a front steering turning vehicle (to the left) at low speed where the tire lateral forces are low and they roll without slipping. Assuming small angles, the left and right wheels steering angles can be estimated by:

$$\delta_r = \frac{L}{R + \frac{l}{2}} \quad (10.3)$$



■ Fig. 10.3
Ackermann angle

$$\delta_l = \frac{L}{R - \frac{t}{2}} \quad (10.4)$$

Where t is the tread and L is the track width of the vehicle. Average of the front wheels steering angles using small angles assumption is called Ackermann steering angle ($\delta = \frac{\delta_l + \delta_r}{2} \cong \frac{L}{R}$).

At high speeds and steering angles, the steering angle no longer equals the Ackermann angle (Gillespie 1992), and tire slips should be considered in the estimation:

$$\delta = \frac{180}{\pi} \cdot \frac{L}{R} + \alpha_f - \alpha_r = 57.3 \frac{L}{R} + \frac{1}{C_{\alpha_f}} \cdot \frac{W_f}{g} \cdot \frac{V^2}{R} - \frac{1}{C_{\alpha_r}} \cdot \frac{W_r}{g} \cdot \frac{V^2}{R} \quad (10.5)$$

where W_f is the normal load on the front tires, and W_r is the normal load on rear tires, V is the vehicle speed, R is the turning radius, and C_{α_f} and C_{α_r} are the cornering stiffness of front and rear tires. Parameters W_f and W_r can be estimated by ($W_f = m l_f / L$) and ($W_r = m l_r / L$) when longitudinal acceleration is zero. l_f and l_r are the distances of front and rear axles from the center of gravity of the vehicle (c.g.) as shown in ► Fig. 10.4. With $a_y = \frac{V^2}{Rg}$ in g s, the above equation can be written as:

$$\delta = 57.3 \frac{L}{R} + K a_y \quad (10.6)$$

K (deg/g) is known as understeer gradient (Gillespie 1992; Rajamani 2006).

If $K = 0$: the vehicle is turning in Neutral steer, which means that the turning radius is constant and does not change with increasing speed.

If $K > 0$: the vehicle is understeering and the turning radius changes linearly with increasing lateral acceleration. The front tires steering angle should increase with increasing speed to maintain the curve.

If $K < 0$: the vehicle is oversteering and the steering angle should be reduced to maintain the curve. It is clear that one cannot reduce the steering angle below zero, so the velocity at which $\left(57.3 \frac{L}{R} = K \frac{V^2}{Rg}\right)$ is called critical speed and the vehicle becomes directionally unstable beyond this speed.


Other factors that affect steering angle are traction forces, aligning torque, roll steer, and camber angles, lateral load transfer, and the steering system. The total value of understeer gradient can be derived by summing these effects (Gillespie 1992).

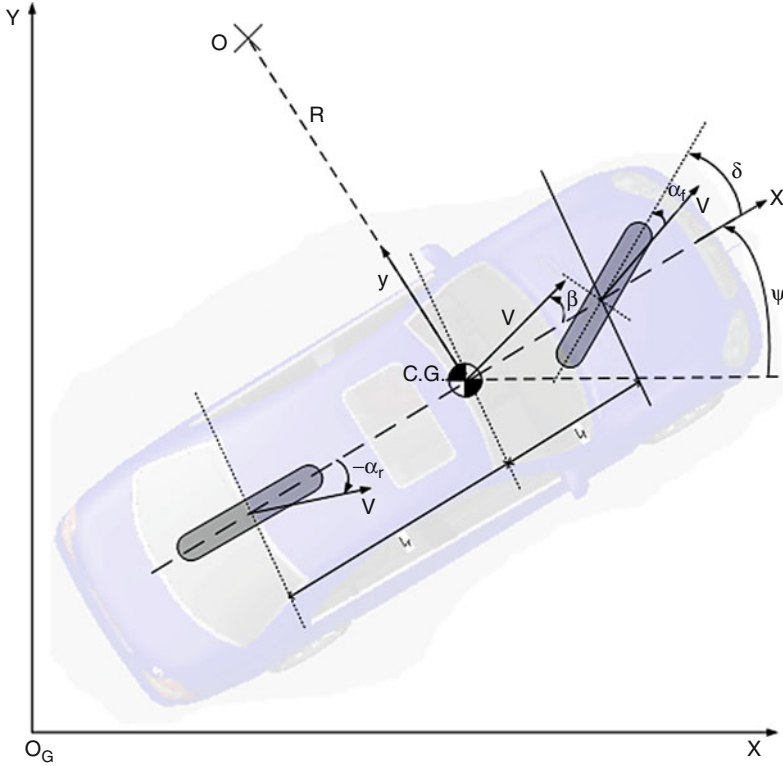
2.4 Suspension System

The mechanism that connects the wheels to the vehicle body is the suspension. Suspension systems can be divided to two major categories: (1) solid axles, and (2) independent axles. Solid axle is the mechanism in which the wheels are connected using a solid axle and move together. This mechanism lets the body to move up and down and roll but does not let other relative movements of the body relative to the wheels. Independent suspension let the wheels move vertically without affecting the opposite wheel. Independent suspension systems are usually more expensive but are the preferred system for most of the passenger cars. In the single track model of the vehicle (i.e., bicycle model) with small lateral acceleration, the effect of suspension is usually ignored.

3 Vehicle Model

The most common model for lateral dynamics of the vehicle is a single track model of the vehicle known as bicycle model (Rajamani 2006; Milliken and Milliken 1995). Comparison of a complex six degrees of freedom (6 DOF) model and a simple (2 DOF) model of the vehicle has shown that the response of both models remain close to each other in the absence of abrupt steering input (Peng and Tomizuka 1990). Bicycle model can be extended for longer wheelbases (such as a city bus) by having two masses to model the vehicle to address the actual mass distributions (Peng and Tomizuka 1990; Ackermann 1994). In terms of four-wheel steering versus two wheel steering, Hiraoka et al. 2009, used vehicle dynamics model of CarSim for two and four wheel steering and concluded that 4-WS vehicle resulted in lower rear path deviation and side slip angle (at constant radius) than 2-WS system (Hiraoka et al. 2009). However, it should be noted that the achievable benefit from four-wheel steering is not significant for an experienced driver (Lee 1995), and front wheel input provides most of the benefits of four-wheel steering (Alleyne 1997).

The bicycle model of a front steering vehicle is depicted in  Fig. 10.4. In deriving a simplified dynamic model, aerodynamic gusts and road irregularities are not considered,



■ Fig. 10.4

Bicycle Model, Reprinted with permission from ASME paper (Soudbakhsh et al. 2010)

and the forward velocity is assumed to be constant. All primary control and disturbance forces, except aerodynamic forces are generated in the tire-road contact patches, which are in general due to lateral slip and lateral inclination of the tire. In the model shown in Fig. 10.4, the two degrees of freedom are vehicle yaw angle (ψ), and vehicle lateral position (y); point O is the center of rotation of the vehicle, and y is measured with respect to it, and yaw is measured based on the global coordinate XY with respect to X axis (Rajamani 2006).

From applying Newton's second law along y-axis:

$$\Sigma F_x : ma_y = F_{yf} + F_{yr} + F_{bank} \quad (10.7)$$

Where $a_y = \left(\frac{d^2 y}{dx^2}\right)$ is lateral acceleration in inertial coordinate, F_{yf} , and F_{yr} are the tire lateral forces and F_{bank} is the force produced as a result of banking angle (ϕ), and can be estimated by $F_{bank} = mg \sin(\phi)$.

Momentum about z axis is equal to $I_z \dot{r}$, so:

$$I_z \dot{r} = l_f F_{yf} - l_r F_{yr} \quad (10.8)$$

Where l_f and l_r are distances of the front and rear axles from center of gravity of the vehicle, respectively, and r is the yaw rate of the vehicle $\left(r = \dot{\psi} = \frac{d\psi}{dt}\right)$.

Slip angles are derived using the geometrical relation shown in ► Fig. 10.4:

$$V \sin \beta + l_f \dot{\psi} = V_x \tan(\delta - \alpha_f) \quad (10.9)$$

$$V \sin \beta - l_r \dot{\psi} = V_x \tan(-\alpha_r) \quad (10.10)$$

$$V \sin \beta = V_y = \dot{y} \quad (10.11)$$

Using the above equations and assumption of small angles yields the following relations for cornering angles:

$$\delta - \alpha_f = \frac{V_y + l_f \dot{\psi}}{V_x} \quad (10.12)$$

$$\alpha_r = \frac{V_y - l_r \dot{\psi}}{V_x} \quad (10.13)$$

In addition to ► Eqs. 10.1–10.13, the following kinematic relations are used to compute the position of the vehicle:

$$\dot{Y} = V_x \sin(\psi) + \dot{y} \cos(\psi) \quad (10.14)$$

$$\dot{\psi} = r \quad (10.15)$$

$$\dot{X} = V_x \cos(\psi) - \dot{y} \sin(\psi) \quad (10.16)$$

3.1 Tire Models

Many tire models have been introduced and used by researchers; however, besides linear estimation of tire response, magic formula tire model has been widely used. The following section includes a brief description and formulation of the latest magic formula tire model (Pacejka 2006). In addition, an introduction to brush tire model and SWIFT tire model are also included. At the end of this part, a linear estimation of tire response based on magic formula tire model is presented. Other examples of tire models that have been used by researchers but are not included in this chapter are Doguff tire model which is an analytical tire model with possibility of modeling both longitudinal and lateral tire forces (Dugoff et al. 1970) and LuGre dynamics friction model (Alvarez et al. 2005).

3.1.1 Magic Formula Tire Model

Magic formula tire model (Pacejka 2006; Pacejka and Bakker 1992) is a nonlinear and semiempirical tire model, which is able to simulate the generated forces in the tires. Using

this formulation, the lateral forces generated in the tires under pure slip condition can be estimated by the following equation:

$$F_y = D_y \sin(C_y \tan^{-1}(B_y x_y - E_y (B_y x_y - \tan^{-1}(B_y x_y)))) + S_{vy} \quad (10.17)$$

where, F_y is the lateral force generated by the tire. Magic formula tire model in the pure slip condition as given by Pacejka (2006) is presented below:

$$\mu_y = (P_{Dy1} + P_{Dy2} df_z) \cdot (1 - p_{Dy3} \gamma^{*2}) \cdot \frac{\lambda_{\mu y}}{1 + \frac{\lambda_{\mu V} V_z}{V_0}} \quad (10.18)$$

$$D_y = \mu_y F_z \zeta_2 \quad (10.19)$$

$$C_y = P_{Cy1} \cdot \lambda_{Cy} \quad (> 0) \quad (10.20)$$

$$K_{ya0} = p_{Ky1} F'_{z0} \sin \left[2 \arctan \left\{ \frac{F_z}{p_{Ky2} F'_{z0}} \right\} \right] \cdot \lambda_{Ey} \quad (10.21)$$

$$K_{ya} = K_{ya0} \cdot (1 - p_{Ky3} \gamma^{*2}) \cdot \zeta_3 \quad (10.22)$$

$$B_y = \frac{K_{ya}}{C_y D_y + \varepsilon_y} \quad (10.23)$$

$$S_{Hy} = (p_{Hy1} + p_{Hy2} df_z) \cdot \lambda_{Hy} + p_{Hy3} \gamma^* \cdot \lambda_{Ky\gamma} \zeta_0 + \zeta_4 - 1 \quad (10.24)$$

$$S_{Vy} = F_z \{ (p_{Vy1} + p_{Vy2} df_z) \lambda_{Vy} + (p_{Vy3} + p_{Vy4} df_z) \gamma^* \lambda_{Ky\gamma} \} \lambda'_{\mu y} \zeta_2 \quad (10.25)$$

$$E_y = (p_{Ey1} + p_{Ey2} df_z) \cdot \{ 1 - ([E_{y3} + p_{Ey4} \gamma^*] \operatorname{sgn}(\alpha_y)) \} \cdot \lambda_{Ey} \quad (10.26)$$

$$\alpha_y = \alpha^* + S_{Hy} \quad (10.27)$$

$$F_y = D_y \sin[C_y \{ B_y \alpha_y - E_y (B_y \alpha_y - \arctan(B_y \alpha_y)) \}] + S_{Vy} \quad (10.28)$$

where, λ_i s are scaling factors, F_y is the lateral load, and α and γ are the slip and camber angles, respectively. S_{Hy} and S_{Vy} are the horizontal and vertical shifts in α and F_y , respectively. B is the stiffness factor, C is the shape factor, D is the peak value and E is the curvature factor.

3.1.2 Brush Model

In this model, tire is considered as a row of elastic bristles that deflect in a direction parallel to the road surface (Pacejka 2006). Side slip occurs when the velocity make an angle with the wheel's plane. On the other hand, if the tire effective rolling velocity multiplied by the rotational velocity of the tires differs from the forward velocity, fore and aft slip occurs.

3.1.3 Swift Tire Model

SWIFT tire model (Short Wavelength Intermediate Frequency Tyre) model is a semiempirical tire model based on a rigid ring type of tire. This tire model was developed by joint cooperation of TNO and Delft University to model the tire behavior higher frequencies (up to 60 Hz) and short road obstacles. This model can be used for active chassis control system development, and suspension vibration analysis (Pacejka 2006).

3.1.4 Linear Tire Model

For small slip angles, the lateral forces generated on the tires (F_{yf} and F_{yr}) can be estimated using a linearized version of the magic formula tire model, and represented by the following equations:

$$F_{yf} = C_{\alpha f} \alpha_f \quad (10.29)$$

$$F_{yr} = C_{\alpha r} \alpha_r \quad (10.30)$$

Where, α_f and α_r are the front and rear tire slip angles, respectively. $C_{\alpha f}$ ($C_{\alpha r}$) is the cornering stiffness of the front (rear) wheels and can be estimated by the initial slope of the magic formula tire model in the pure slip condition (Pacejka 2006) as in the following relation:

$$C_a = \left. \frac{\partial F_y}{\partial \alpha} \right|_{\alpha=\gamma=0} = p_{Ky1} F'_{z0} \sin \left[2 \arctan \left\{ \frac{F_z}{p_{Ky2} F'_{z0}} \right\} \right] \cdot \lambda_{Ey} \quad (10.31)$$

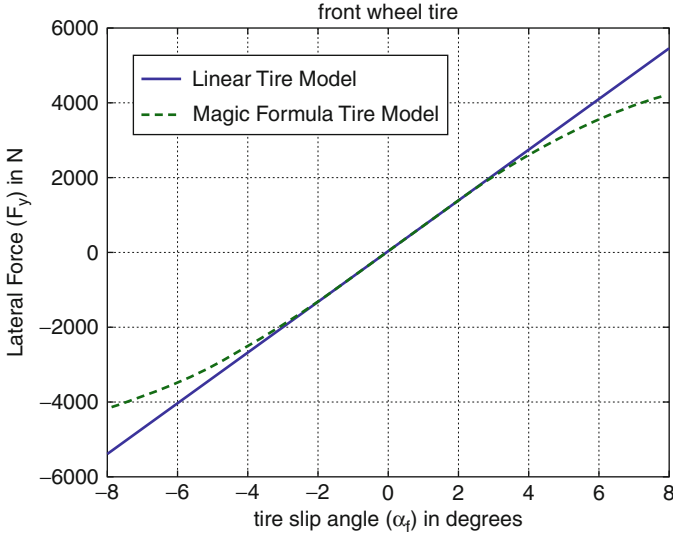
Where, p_{Ky1} , p_{Ky2} , and λ_{Ey} are constant parameters, F'_{z0} is the adapted nominal load (Pacejka 2006), γ is the camber angle, and F_z is the vertical load on the tire.

► Figure 10.5 shows the comparison between the estimated lateral forces generated in front tire using linear tire model, and magic formula tire model.

Using linear tire model with the assumption of small slip angles, and ignoring the bank angle, the dynamics of the system will have the following form:

$$\frac{d}{dt} \begin{bmatrix} \dot{y} \\ r \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} \dot{y} \\ r \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \delta \quad (10.32)$$

Where $a_1 = -\frac{2C_{\alpha f} + 2C_{\alpha r}}{mV_x}$, $a_2 = -\left(V_x + \frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{mV_x}\right)$, $a_3 = \frac{-2C_{\alpha f}l_f + 2C_{\alpha r}l_r}{I_z V_x}$,
 $a_4 = -\frac{2C_{\alpha f}l_f^2 + 2C_{\alpha r}l_r^2}{I_z V_x}$, $b_1 = \frac{2C_{\alpha f}}{m}$ and $b_2 = \frac{2C_{\alpha f}l_f}{I_z}$.



■ Fig. 10.5

Comparison of magic formula tire model (Pacejka 2006) and linear tire model for 205/60R15 tires with normal load of about 4730 N on the tire

Combining the above equation with the kinematic relations (► 10.14–10.16) results in the following state-space system with lateral position, lateral velocity, yaw angle, and yaw rate as the state variables (► 10.33):

$$\frac{d}{dt} \begin{bmatrix} y \\ \dot{y} \\ \psi \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{\alpha f} + 2C_{\alpha r}}{mV_x} & 0 & -\left(V_x + \frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{mV_x}\right) \\ 0 & 0 & 0 & 1 \\ 0 & \frac{-2C_{\alpha f}l_f + 2C_{\alpha r}l_r}{I_z V_x} & 0 & -\frac{2C_{\alpha f}l_f^2 + 2C_{\alpha r}l_r^2}{I_z V_x} \end{bmatrix} \begin{bmatrix} y \\ \dot{y} \\ \psi \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2C_{\alpha f}}{m} \\ 0 \\ \frac{2C_{\alpha f}l_f}{I_z} \end{bmatrix} \delta + \begin{bmatrix} 0 \\ g \\ 0 \\ 0 \end{bmatrix} \sin(\phi) \quad (10.33)$$

In ► Eq. 10.33, $C_{\alpha f}$ and $C_{\alpha r}$ are cornering stiffness of each of the front and rear tires, respectively. ϕ is the bank angle, which is zero for straight roads, and δ is the steering angle.

3.2 Error Coordinates

The system of ► Eq. 10.27 is not controllable; so in order to design a controller such as linear quadratic regulator, a new coordinate system based on the errors from the desired path can be used. The position error (e_1) was defined as the distance of the center of gravity (c.g.) of the vehicle from the desired trajectory, and the yaw error (e_2) was defined as the difference between the actual orientation of the vehicle and the desired yaw angle

(Rajamani 2006; Soudbakhsh and Eskandarian 2011). The relations to calculate the error vectors are given in the following equations:

$$\ddot{e}_1 = a_y - a_{y_{des}} = (\ddot{y} + V_x \dot{\psi}) - a_{y_{des}} \quad (10.34)$$

$$e_2 = \psi - \psi_{des} \quad (10.35)$$

In the above equations, $(\cdot)_{des}$ represents the ideal value on the desired trajectory, and a_y is the lateral acceleration of the vehicle.

As an example if the objective is to keep the vehicle in its lane with curvature of $\kappa = 1/R$; the desired values assuming constant longitudinal velocity will be $\dot{\psi}_{des} = V_x/R$, and $a_{y_{des}} = V_x^2/R$, and \blacklozenge Eq. 10.33 can be represented by:

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} e_1 \\ \dot{e}_1 \\ e_2 \\ \dot{e}_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ \frac{2C_{\alpha f}}{m} \\ 0 \\ \frac{2C_{\alpha f} l_f}{I_z} \end{bmatrix} \delta + \begin{bmatrix} 0 \\ -\left(V_x + \frac{2C_{\alpha f} l_f - 2C_{\alpha r} l_r}{mV_x}\right) \\ 0 \\ -\frac{2C_{\alpha f} l_f^2 + 2C_{\alpha r} l_r^2}{I_z V_x} \end{bmatrix} \dot{\psi}_{des} + \begin{bmatrix} 0 \\ g \\ 0 \\ 0 \end{bmatrix} \sin(\phi) \\ &+ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{\alpha f} + 2C_{\alpha r}}{mV_x} & \frac{2C_{\alpha f} + 2C_{\alpha r}}{m} & \frac{-2C_{\alpha f} l_f + 2C_{\alpha r} l_r}{mV_x} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{-2C_{\alpha f} l_f + 2C_{\alpha r} l_r}{I_z V_x} & \frac{2C_{\alpha f} l_f - 2C_{\alpha r} l_r}{I_z} & -\frac{2C_{\alpha f} l_f^2 + 2C_{\alpha r} l_r^2}{I_z V_x} \end{bmatrix} \begin{bmatrix} e_1 \\ \dot{e}_1 \\ e_2 \\ \dot{e}_2 \end{bmatrix} \end{aligned} \quad (10.36)$$

3.3 Global Position of the Vehicle

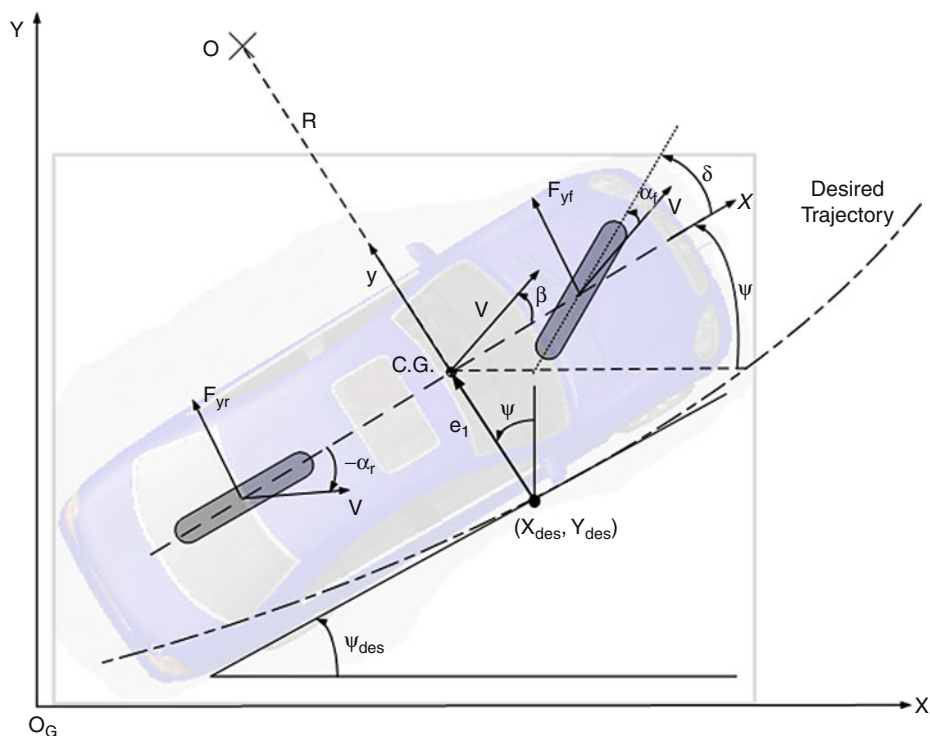
Location of the vehicle, using error coordinate system values (e_1, e_2) , and desired location of the vehicle (X_{des}, Y_{des}) as in \blacklozenge Fig. 10.6 can be obtained by the following transformation:

$$\begin{cases} X = X_{des} - e_1 \sin(\psi) \\ Y = Y_{des} + e_1 \cos(\psi) \end{cases} \quad (10.37)$$

3.4 Using Yaw Rate and Slip Angle as State Space Variables

Body side slip angle (β) is defined as the angle between vehicle longitudinal axis, and direction of travel (\blacklozenge Fig. 10.4), and it can be estimated by $\beta = \dot{y}/V_x$ under small angle assumptions. Body side slip angle and yaw rate ($r = \dot{\psi}$) have been used by some researchers as the state variables (Ackermann et al. 1995; Guldner et al. 1994; Guvenc and Guvenc 2002), and the vehicle lateral dynamics can be described as in \blacklozenge Eq. 10.38:

$$\begin{cases} mV_x \left(\frac{d\beta}{dt} + r \right) = F_{yf} + F_{yr} + F_{bank} \\ I_z \dot{r} = l_f F_{yf} - l_r F_{yr} \end{cases} \quad (10.38)$$



■ Fig. 10.6

Global and error coordinates, reprinted with permission from SAE paper 2010-01-0459

© 2010 SAE International

Having small angle assumptions, \bullet Eq. 10.38 yields to the following set of equations:

$$\begin{cases} \frac{d\beta}{dt} = -r + \frac{C_{\alpha f}}{mV_x} \left(\delta - \beta - \frac{l_f r}{V_x} \right) + \frac{C_{\alpha r}}{mV_x} \left(-\beta + \frac{l_r r}{V_x} \right) + \frac{g \sin(\phi)}{V_x} \\ \frac{dr}{dt} = \frac{l_f C_{\alpha f}}{I_z} \left(\delta - \beta - \frac{l_f r}{V_x} \right) - \frac{l_r C_{\alpha r}}{I_z} \left(-\beta + \frac{l_r r}{V_x} \right) \end{cases} \quad (10.39)$$

Or in the state space form:

$$\begin{bmatrix} \dot{\beta} \\ \dot{r} \end{bmatrix} = \begin{bmatrix} \frac{-C_{\alpha f} - C_{\alpha r}}{mV_x} & -1 - \frac{C_{\alpha f} l_f - C_{\alpha r} l_r}{mV_x^2} \\ \frac{-l_f C_{\alpha f} + l_r C_{\alpha r}}{I_z} & \frac{-C_{\alpha f} l_f^2 - C_{\alpha r} l_r^2}{I_z V_x} \end{bmatrix} \begin{bmatrix} \beta \\ r \end{bmatrix} + \begin{bmatrix} \frac{C_{\alpha f}}{mV_x} \\ \frac{l_f C_{\alpha f}}{I_z} \end{bmatrix} \delta + \begin{bmatrix} \frac{g}{V_x} \\ 0 \end{bmatrix} \sin \phi \quad (10.40)$$

Steady state steering angle (δ_{ss}) can be derived by setting \dot{r} , and $\ddot{\psi}$ to zero, and having $r = v_x/R$:

$$\delta_{ss} = \frac{L}{R} + \frac{v^2}{R} \left(\frac{ml_f}{C_{\alpha r} L} - \frac{ml_r}{C_{\alpha f} L} \right) = \frac{L}{R} + K_v a_y \quad (10.41)$$

Where, K_y is the understeer gradient. Also, global position of the vehicle can be calculated from the following set of equations:

$$\begin{cases} \dot{X} = V\cos(\psi + \beta) \\ \dot{Y} = V\sin(\psi + \beta) \end{cases} \quad (10.42)$$

4 Controllers

Many controllers have been developed to perform successful lane change or lane-keeping maneuvers. Simple proportional-plus-Derivative (PD) controllers cannot result in both satisfactory levels of response and convergence; however, using two sets of PD like controllers tuned with linear quadratic (LQ) control theory provides better performance (Mouri and Furusho 1997). Furusho and Mouri (1999), from Nissan Motor Company research center, used linear quadratic control to automatically keep the vehicle in curved lane (Furusho and Mouri 1999). Chen and Tan (1998), used dynamic, frequency dependent distance $d(s)$ instead of static look ahead distance to decouple yaw motion from the lateral dynamics of the vehicle (Chen and Tan 1998). Guldner et al. 1999, augmented look-down reference system with an additional sensor at the tail bumper of the car to overcome speed restrictions of these systems (Guldner et al. 1999). Chan and Tan 2001, used a simple bicycle model, and a lag-lead compensator for a look forward controller to study the feasibility of using active steering for postcrash scenarios, and found their approach to be effective and a promising strategy for postcrash handling of the car (Chan and Tan 2001).

Guldner et al. (1994), used sliding mode controller for lane change maneuver on a simple bicycle model (Guldner et al. 1994). Edwards et al. (2005) used sliding mode observers to detect situations such as oversteering and understeering (Edwards et al. 2005). They used bicycle model and Doguff tire model, and used lateral acceleration and yaw rate as state variables. They stated that the tradeoff between detection time and sensors noise attenuation is very important in designing such observers (Edwards et al. 2005). Schorn et al. 2006, used Extended Kalman filter to design a sideslip angle feedback controller to follow a desired trajectory (Schorn et al. 2006). Akar and Kalkkuhl 2008, developed a sliding mode controller for a four-wheel steering vehicle with a bicycle model and showed that this method can be used for a variety of vehicles (Akar and Kalkkuhl 2008). Following success of electronic stability control system in improving the performance of average drivers, there are ongoing projects in improving the ability of the controllers in assisting drivers with various skills in lane-keeping and other lateral assistance tasks by trying to use the tires near or even at their saturated limit by using appropriate steering, brake, and throttle inputs. This type of stability at the limits is inspired by performance of race car drivers and an example of it can be found in Talvala et al. (2011).

In the recent years, following successful implementation of model predictive controller (MPC) in several industrial projects (Qin and Badgwell 2003), interest to use MPC in lateral control of the vehicle has grown in the automotive research centers. For example, Keviczky et al. (2006), simulated bicycle model of the vehicle with a model predictive

controller to autonomously perform a double lane change maneuver and concluded that MPC policy resulted in a disturbance rejection of 10 m/s and vehicle speed of 17 m/s (Keviczky et al. 2006). In 2007, researchers at Ford Motor Company tested a predictive approach for an autonomous vehicle to perform a double lane change maneuver on an icy track in which they tested three types of model predictive controllers (Falcone et al. 2007). In 2009, Lee and Yoo designed a model predictive controller using error coordinates of the bicycle model of the vehicle for trajectory following with velocities as high as 80 km/h (Lee and Yoo 2009). In 2010, Anderson et al. developed a threat assessment algorithm using model predictive control and a simple bicycle model to keep the vehicle in safe corridor, while performing a collision avoidance maneuver (Anderson et al. 2010).

Use of adaptive control methods to perform lane change maneuver has been investigated too. Kehtarnavaz and Sohn (1991) used neural network to emulate human driving. They compared backpropagation and functional-link networks in terms of training and recall capabilities, and concluded that functional-linked network provides better results (Kehtarnavaz and Sohn 1991). Funabiki and Mino 1993 used back propagation to develop a neural-network based steering control system (Funabiki and Mino 1994). Hessburg and Tomizuka (1994) developed a fuzzy logic controller to use preview information of the road, and appropriate gains with regard to the velocity of the vehicle were chosen with a gain scheduling rule base. Tests on the vehicle showed the robustness of the controller and its ability to handle large number of input variables (Hessburg and Tomizuka 1994). Nagai et al. (1995) used neural network to design a four wheel steering vehicle lateral control system by modeling nonlinear response of tire models and using combination of linear tire model and nonlinear neural network. Testing of the system with a full vehicle dynamics model showed that it can improve the handling and stability of the vehicle (Nagai et al. 1995). Nagai et al. (1997), used genetic algorithm to analyze braking and handling inputs to perform an obstacle avoidance maneuver, and concluded that high steering gain and short preview time are required for performing a successful maneuver (Nagai et al. 1997). Naranjo et al. (2008) developed a system to perform double lane change maneuver by using fuzzy controllers and mimicking human drivers behavior (Naranjo et al. 2008).

It is seen that tremendous research has gone into designing of vehicle lateral controllers because the problem is very challenging, especially at extreme speeds or trajectories (high steering angle and rates) during evasive maneuvers at which the tire and vehicle dynamics become highly nonlinear. External disturbances add to the control challenge as well.

5 Desired Trajectory for Lane-Change and Lane Keeping Maneuvers

Lane change and lane-keeping maneuvers are two of the major goals of vehicle lateral control. These maneuvers are relatively slow and can be done by considering occupants comfort zone. Lane change maneuvers usually follow a desired trajectory. Many

trajectories have been proposed by researchers and the following section, gives a description and comparison of two of the most famous ones.

In Sledge and Marshek (1997), different analytical trajectories were compared to each other and concluded that the fifth order polynomial is the preferred candidate for this kind of maneuver. A polynomial of degree 5 was defined in (Nelson 1989) by assuming a zero lateral velocity and acceleration at both ends and reaching a final position (x_e, y_e) at the end of the maneuver. This polynomial has the following form (► 10.43):

$$y(x) = y_e \left(10 \left(\frac{x}{x_e} \right)^3 - 15 \left(\frac{x}{x_e} \right)^4 + 6 \left(\frac{x}{x_e} \right)^5 \right) \quad (10.43)$$

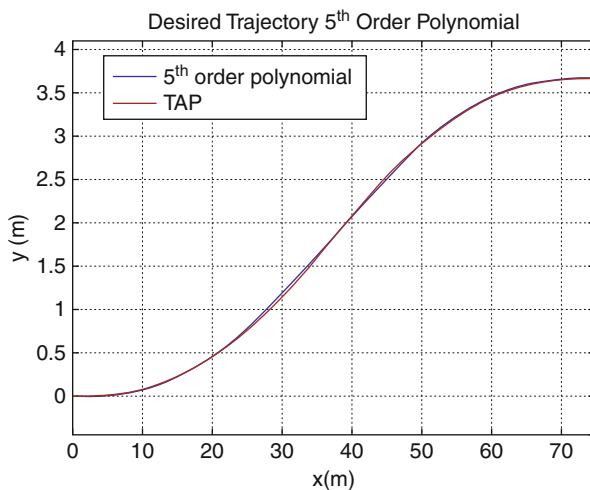
An alternative approach to define the desired trajectory is to start from acceleration profile and deriving the trajectory by double integration. By assuming an upper bound for lateral acceleration and its rate of change (jerk/jolt), a trapezoidal acceleration profile (TAP) is defined by (► 10.44) (Chee and Tomizuka 1994)

$$a_y = \begin{cases} J_{\max} t & 0 \leq t < t_1 \\ J_{\max} t_1 & t_1 \leq t < t_2 \\ -J_{\max} t + J_{\max}(t_1 + t_2) & t_2 \leq t < t_3 \\ -J_{\max} t_1 & t_3 \leq t < t_4 \\ J_{\max}(t - 2t_2 - t_1) - 4J_{\max} t_1 & t_4 \leq t \leq t_5 \end{cases} \quad (10.44)$$

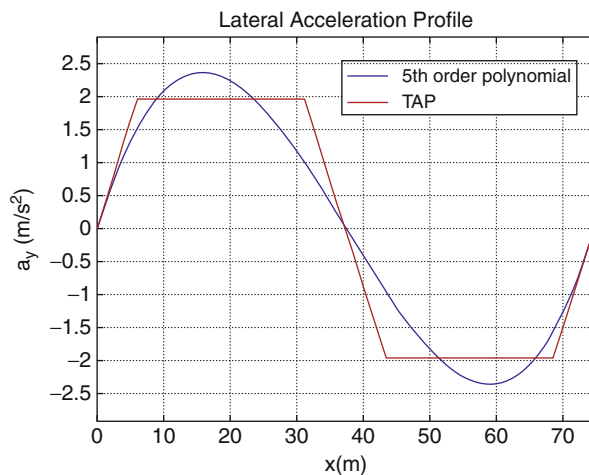
► Figure 10.7 shows desired lane change trajectories generated by fifth order polynomial and trapezoidal acceleration profile with final time of about 3 s and velocity of 25 m/s. acceleration profile of these trajectories are shown in ► Fig. 10.8. As shown in ► Fig. 10.8, if the jerk input of trapezoidal acceleration profile (TAP) is matched to the initial jerk of the fifth order polynomial, for the same desired final lateral displacement, the resultant acceleration profile has a lower peak than the fifth order polynomial. Consequently, with the same available maximum lateral acceleration (due to the coefficient of friction between tire and road), TAP can accommodate more severe maneuvers.

However, it should be noted that for performing severe (evasive) maneuvers such as collision avoidance lane changes, the lateral acceleration of the vehicle can become very high using TAP or fifth order polynomial and result in different responses of the tires (i.e., nonlinear region) and other vehicle components with high slip angles (Soudbakhsh and Eskandarian 2011; Soudbakhsh et al. 2010). For this reason, many different trajectory planning approaches have been developed to perform these kinds of severe maneuvers, for example, by using minimum lateral acceleration (Soudbakhsh et al. 2010), or minimum front slip angles (Anderson et al. 2010). Besides the traditional optimization approaches, different algorithms for finding optimal trajectories using search methods such as rapidly-exploring random tree algorithm (Kuwata et al. 2009) have been suggested by some researchers for vehicle motion planning.

These methods can result in a much lower lateral acceleration profile and provide better stability. Since another chapter of the book is dedicated to these kinds of maneuvers, it will not be discussed any further here.



■ Fig. 10.7
Fifth order polynomial trajectory versus TAP trajectory



■ Fig. 10.8
Comparison of the trapezoidal acceleration profile and the fifth order polynomial

6 Some Experimental Studies

Maeda et al. (1977) tested 20 male subjects on a test track to categorize their response. They concluded that the drivers regardless of their experience steer the vehicle with the same rate, and this maximum steering rate concentrated in the range of 700–900 deg/s

(Maeda et al. 1977). They showed that inexperienced drivers have problem with performing a stable maneuver rather than initializing the maneuver (Maeda et al. 1977). The obstacle in their study was a foam object shaped-like back of a car which suddenly jumped into the driver's lane from behind a partition.

Yoo et al. (1996) tested 24 subjects in a car driving simulator and showed that visual icons are more effective than showing text to alert subjects of an upcoming obstacle on the road (Yoo et al. 1996). In their study, they could not test each subject more than one time since subject slowed down for the second encounter, so they compared results of the first test only (some with text and some with icon) (Yoo et al. 1996).

Chee (1997) used a virtual desired trajectory to control lateral motion of the vehicle using a bicycle model. He used a reduced order Kalman filter to estimate immeasurable parameters in his study, which were lateral position and lateral velocity (Chee 1997). This project was done on a dedicated lane to perform automatic lane change maneuver.

Shin et al. (2003) added tire nonlinearity, variable gear ratio, and lateral load transfer to the bicycle model, and used measured lateral acceleration at the center of gravity of the vehicle to estimate yaw rate based on Kalman filter (Shin et al. 2003). Their estimations of yaw rate were different by less than 2 deg/s from experimental results (Shin et al. 2003).

In the recent years there has been extensive research on the autonomous and semiautonomous vehicles. Examples of the experimental studies in the recent years for vehicle lateral control are given in Anderson et al. (2010). and Yih and Gerdes (2005). DARPA Grand Challenge for autonomous driving has motivated many research groups in developing driverless vehicles and many systems and algorithms have been demonstrated successfully in these projects. Some details about different components and algorithms used in these vehicles are presented in references (Leonard et al. 2009; Kuwata et al. 2009; Urmson et al. 2008).

7 Case Studies

Many automotive OEMs developed their own lane-keeping system and a few are commercially available. There are some aspects of these systems that are not related to the dynamics of the vehicle but are equally important. For example, lane marker detection, including their shape and position, is one of the prerequisites of lane departure warning or lane-keeping systems. Lane markers' detection is usually done by evaluating the contrast between the road and the lane markings using onboard cameras. Although there are many lane departure warning and lane-keeping systems available today, the lane change systems are mostly demonstrated but are not in the market yet. The required situational awareness and higher level decision making to execute an evasive maneuver in the presence of other traffic still requires more research and cautious considerations.

Some of the commercially available lane-keeping systems is listed herein. This is by no means a comprehensive coverage of all available systems, nor does it imply any advantage of the listed OEMs over those not covered here. These are presented merely as examples of many other available systems, highlighting typical issues and concerns. Although some of these systems may control the steering subtly, none of these are complete steering control

for lane change or severe maneuvers. The experimental case studies of steering control are covered in the ● Sect. 6 of this chapter.

Mercedes-Benz has developed a lane-keeping system to reduce number of lane departure accidents. While most of the lane departure systems monitor the left/right signals as a measure to activate the system or not, this system monitors more parameters such as braking or acceleration before the lane departure to activate the system. The system is activated if it predicts an unintentional lane departure by vibrating the steering wheel and applying counter steering at the same time to bring the vehicle back to its original lane (Danielson 2008).

Ford motor company's Lane-Keeping Aid detects white lane markings. This system applies a gentle counter-steering torque to correct the vehicle's position in the lane. However, if the driver does not take appropriate action to correct the vehicle position and the vehicle crosses its lane the steering wheel vibrates to warn the driver (www.euroncap.com).

DENSO Corporation developed a steering assist electronic control unit (ECU). This system calculates a target steering torque based on the data from the vision sensor, and then sends a steering torque signal to an electric power steering (EPS) ECU to control the EPS motor (DENSO).

Toyota Lane-Keeping Assist (such as the ones on the Lexus LS 460 and LS 460 AWD) uses a stereo imaging camera to monitor white line road markings. Besides lane change warning, this system helps the driver by applying steering torque to keep the vehicle in its lane (Lexus).

Continental has a Lane Departure Warning (LDW) driver assistance system to alerts the driver about their possible lane departure. Their Lane-Keeping System (LKS) also apply steering torque.

The TRW system is also using the information from the video sensor. This system is activated if the driver crosses a road lane marking or the edge of the road; the system provides a counter-steering torque to help guide the driver back to the center of the lane. The TRW system brings the vehicle to the center of the lane if the driver does not take control of the vehicle from the system (TRW).

Many other companies also have some kind of lane-keeping system. There are still technical issues with detecting the lane and true intention of the driver. Lane detection may fail when lane markings are missing or there are multiple lane markings in work zones. Another problem is that many of the new technologies work only in a range of velocities. This can be an issue when the driver expects the system's intervention but the system does activate below or above a certain speed.

8 Electronic Stability Control (ESC)

Electronic Stability Control (ESC) is one of the major safety components achieved by analyzing vehicle lateral dynamics. This system is a closed loop control system integrated in the brake and drive train systems of the vehicle and has many trade names such as Electronic

Stability Program (ESP), Dynamic Stability Control (DSC), Vehicle Dynamic Control (VDC), Vehicle Stability Assist (VSA), and Vehicle Stability Control (VSC). According to NHTSA (adopted based on Report J2564 of the Society of Automotive Engineers (SAE)), ESC is defined as a closed loop control system to limit oversteer and understeer of the vehicle that works for all range of vehicle speed above a certain threshold that (1) Augment vehicle directional stability by correcting vehicle yaw torques through applying and adjusting the vehicle brakes individually (2) able to determine yaw rate, its sideslip angle, and monitor driver steering input. Also, ESC system should be able to control braking of all wheels of the vehicle individually, and should work with any acceleration of the vehicle and its deceleration even when other systems such as antilock braking system (ABS) or Traction Control systems are activated. Performance of ESC is evaluated by testing a vehicle in performing a maneuver known as 0.7 Hz Sine with Dwell maneuver at speed of 80 kph (50 mph) (NHTSA).

The controller of the ESC system compares the desired motion of the vehicle, which can be derived using steady state response of the vehicle to the applied steering angle at the velocity of the vehicle (using a model similar to the one introduced in [Sect. 3.4](#)), to the actual motion of the vehicle to correct vehicle's path. NHTSA estimates that this system will reduce single vehicle crashes of passenger cars by 34% and single vehicle crashes of sport utility vehicles (SUVs) by 59% and will save more than 5,000 lives annually and many more injuries annually once all the light vehicle are equipped with such a system (NHTSA). An example of the challenges and modeling of the electronic stability control can be found in Tseng et al. (1999), where a relative steering wheel sensor was used to optimize vehicle stability system performance.

9 Conclusions

Vehicle lateral control has many potential benefits especially in reducing number of single vehicle accidents such as run-off-road crashes. Vehicle lateral control systems can help the drivers to change their lanes safely or assist them in parking or even performing evasive maneuvers. Some of the new systems in the vehicles such as lane-keeping and electronic stability control system are already available in the cars, and even there are regulations to require all the vehicles to be equipped with ESC in the United States by 2012. In this chapter an introduction to major components of the vehicle that affect vehicle lateral dynamics are discussed, and different approaches for vehicle lateral control are reviewed. At the end, some of the successful commercially available lane-keeping systems is reviewed and the still remaining challenges with these kinds of systems are discussed.

The lateral control and assist systems such as lane keeping are available only for a select number of vehicle models. A wider implementation in the market will determine their benefits. A main challenge of these systems is to extend the range of velocities within which the system operates. There are still real life challenges in fully implementing complete lateral vehicle control in lane changes or evasive maneuvers. Road and environmental factors and various disturbances influence the smooth control. Better seamless integration of longitudinal and lateral control is needed. Robust and reliable control

methods that handle most or all unexpected situations and ensure the safest outcome in emergencies still require more R&D and evaluation. The possibility of the driver's counter steering in emergency maneuvers and overtaking control issues should also be further addressed not only from a technical but also from a liability and legal perspective, like many other autonomous driver assistance systems.

References

- Ackermann J (1994) Robust decoupling of car steering dynamics with arbitrary mass distribution. In: Proceedings of the American control conference, 1994, vol 2, pp 1964–1968, 29 June–1 July 1994
- Ackermann J, Guldner J, Sienel W, Steinhauser R, Utkin VI (1995) Linear and nonlinear controller design for robust automatic steering. *Control Syst Technol*, IEEE Trans 3(1):132–143
- Akar M, Kalkkuhl JC (2008) Lateral dynamics emulation via a four-wheel steering vehicle. *Veh Syst Dyn: Int J Veh Mech and Mobil* 46(9): 803–829
- Alleyne A (1997) A comparison of alternative obstacle avoidance strategies for vehicle control. *Vehicle Syst Dyn: Int J Vehicle Mech and Mobil* 27(5):371–392
- Alvarez L, Yi J, Horowitz R, Olmos L (2005) Dynamic friction model-based tire-road friction estimation and emergency braking control. *J Dyn Syst, Meas, Control* 127(1):22–32
- Anderson S, Peters S, Pilutti T, Iagnemma K (2010) An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios. *Int J Veh Auton Syst* 8(2):190–216
- Andreas E, Jochen P, Fredrik G, Jonas E (2007) Toward autonomous collision avoidance by steering. *IEEE Trans Intell Transp Syst* 8(1):84–94
- Bogenrieder R, Fehring M, Bachmann R (2009) Pre-Safe® in rear-end collision situations. In: Proceedings of the 21st international technical conference on the enhanced safety of vehicles. National Highway Traffic Safety Administration, Stuttgart, Germany.
- Chan C, Tan H (2001) Feasibility analysis of steering control as a driver-assistance function in collision situations. *IEEE Trans Intell Transp Syst* 2(1):1–9
- Chee W (1997) Unified lateral motion control of vehicles for lane change maneuvers in automated highway systems. University of California at Berkeley
- Chee W, Tomizuka M (1994) Vehicle lane change maneuver in automated highway systems. California PATH research report, UCB-ITS-PRR-94-22
- Chen C, Tan H-S (1998) Steering control of high speed vehicles: dynamic look ahead and yaw rate feedback. In: Decision and control, 1998. Proceedings of the 37th IEEE conference on, vol 1, pp 1025–1030
- Danielson C (2008) November 12, 2008 At 1:10 PM CST-last update, Mercedes-Benz TecDay special feature: lane keeping assist and speed limit assist. http://www.emercedesbenz.com/Nov08/12_001505_Mercedes-Benz_TecDay_Special_Feature_Lane_Keeping_Assist_And_Speed_Limit_Assist.html. Accessed 31 July 2011
- DENSO (2011) Lane keeping assist system. (<http://www.denso-europe.com/Lane-Keeping-Assist-System-1014130000000001.aspx>). Available: <http://www.denso-europe.com/Lane-Keeping-Assist-System-1014130000000001.aspx>. Accessed 31 July 2011
- Dugoff H, Fancher PS, Segel L (1970) An Analysis of Tire Traction Properties and Their Influence on Vehicle Dynamic Performance
- Edwards C, Hebden R, Spurgeon S (2005) Sliding mode observers for vehicle mode detection. *Veh Syst Dyn* 43(11):823–843
- Falcone P, Borrelli F, Asgari J, Tseng HE, Hrovat D (2007) Predictive active steering control for autonomous vehicle systems. *IEEE Trans Control Syst Technol* 15(3):566–580
- FHWA (2010) Roadway Departure Safety ([internal-pdf://roadwaydeparturesafety.fhwa-2720044548/roadwaydeparturesafety.fhwa.pdf](http://roadwaydeparturesafety.fhwa-2720044548/roadwaydeparturesafety.fhwa.pdf)). Accessed on Apr 7 2010

- Funabiki S, Mino M (1994) Neural-network steering control of an automated guided vehicle. *Electr Eng Japan* 114(7):135–143
- Furusho H, Mouri H (1999) Research on automated lane tracking using linear quadratic control:- control procedure for a curved path. *JSAE Rev* 20(3):325–329
- Gillespie TD (1992) *Fundamentals of vehicle dynamics*. Society of Automotive Engineers, Warrendale
- Guldner J, Utkin VI, Ackermann J (1994) A sliding mode control approach to automatic car steering. In: *Proceedings of the American control conference, 1994*, vol 2, pp 1969–1973, 29 June–1 July 1994
- Guldner J, Sienel W, Han-Shue T, Ackermann J, Patwardhan S, Bunte T (1999) Robust automatic steering control for look-down reference systems with front and rear sensors. *IEEE Trans Control Syst Technol* 7(1):2–11
- Guvenc BA, Guvenc L (2002) Robust two degree-of-freedom add-on controller design for automatic steering. *IEEE Trans Control Syst Technol* 10(1):137–148
- Hedrick JK, Tomizuka M, Varaiya P (1994) Control issues in automated highway systems. *Control Syst Mag, IEEE* 14(6):21–32
- Hessburg T, Tomizuka M (1994) Fuzzy logic control for lateral vehicle guidance. *Control Systems Mag, IEEE* 14(4):55–63
- Hiraoka T, Nishihara O, Kumamoto H (2009) Automatic path-tracking controller of a four-wheel steering vehicle. *Veh Syst Dyn* 47(10):1205–1227
- Ishida S, Gayko JE (2004) Development, evaluation and introduction of a lane keeping assistance system. In: *Proceedings of the intelligent vehicles symposium, 2004 IEEE*, pp 943–944, 14–17 June 2004
- Kawazoe H, Murakami T, Sadano O, Suda K, Ono H (2001) Nissan Motor Co., Ltd Development of a Lane-Keeping Support System, SAE 2001 World Congress, March 2001, Detroit, MI, USA, Paper Number: 2001-01-0797
- Kehtarnavaz N, Sohn W (1991) Steering control of autonomous vehicles by neural networks. In: *Proceedings of the American control conference, 1991*, pp 3096–3101, 26–28 June 1991
- Keviczky T, Falcone P, Borrelli F, Asgari J, Hrovat D (2006) Predictive control approach to autonomous vehicle steering. In: *Proceedings of the American control conference, 2006*, 6 pp, 14–16 June 2006
- Kuwata Y, Karaman S, Teo J, Frazzoli E, How JP, Fiore G (2009) Real-time motion planning with applications to autonomous urban driving. *IEEE Trans Control Syst Technol* 17(5):1105–1118
- Lee A (1995) Performance of four-wheel-steering vehicles in lane change maneuvers. SAE World Congress, Detroit, MI, USA pp 161
- Lee J, Yoo W (2009) An improved model-based predictive control of vehicle trajectory by using nonlinear function. *J Mech Sci Technol* 23(4):918–922
- Leonard J, How J, Teller S, Berger M, Campbell S, Fiore G, Fletcher L, Frazzoli E, Huang A, Karaman S, Koch O, Kuwata Y, Moore D, Olson E, Peters S, Teo J, Truax R, Walter M, Barrett D, Epstein A, Maheloni K, Moyer K, Jones T, Buckley R, Antone M, Galejs R, Krishnamurthy S, Williams J (2009) A Perception-Driven Autonomous Urban Vehicle. In: Buehler M, Iagnemma K, Singh S (eds) *The DARPA urban challenge*. Springer, Berlin/Heidelberg, pp 163–230
- Lexus (2011) Lane Keeping Assist. <http://www.lexus.eu/range/Ls/key-features/safety/safety-lane-keeping-assist.aspx>. Accessed 31 July 2011
- Maeda T, Irie N, Hidaka K, Nishimura H (1977) Performance of driver-vehicle system in emergency avoidance. SAE World Congress, Detroit, MI, USA, pp 518
- Milliken WF, Milliken DL (1995) *Race car vehicle dynamics*. SAE International, Warrendale
- Montiglio M, Martini S, Murdocco V (2006) Development of a lane keeping support system for heavy-trucks. London, United Kingdom
- Mouri H, Furusho H (1997) Automatic path tracking using linear quadratic control theory. In: *Proceedings of the intelligent transportation system. ITSC 1997, IEEE conference on*, pp 948–953, 9–12 Nov 1997
- Nagai M, Ueda E, Moran A (1995) Nonlinear design approach to four-wheel-steering systems using neural networks. *Veh Syst Dyn* 24(4):329–342
- Nagai M, Onda M, Katagiri T (1997) Simulation of emergency obstacle avoidance situations using genetic algorithm. *JSAE Rev* 18(2):158–160
- Nagai M, Mouri H, Raksincharoensak P (2002) Vehicle lane-tracking control with steering torque input. *Veh Syst Dyn* 37:267–278

- Naranjo JE, Gonzalez C, Garcia R, de Pedro T (2008) Lane-change fuzzy control in autonomous vehicles for the overtaking Maneuver. *IEEE Trans Intell Transp Syst* 9(3):438–450
- Nelson W (1989) Continuous-curvature paths for autonomous vehicles. In: *Robotics and automation, 1989. Proceedings, IEEE international conference on*, vol 3, pp 1260–1264, 14–19 May 1989
- NHTSA federal motor vehicle safety standards; electronic stability control systems (FMVSS 126), 49 CFR, parts 571 & 585
- Pacejka HB (2006) *Tyre and vehicle dynamics*, 2nd edn. Butterworth-Heinemann, Oxford
- Pacejka H, Bakker E (1992) The magic formula tyre model. *Veh Syst Dyn* 21:1–18
- Peng H, Tomizuka M (1990) Lateral control of front-wheel-steering rubber-tire vehicles. *California PATH*
- Pohl J, Ekmark J (2003) Development of a haptic intervention system for unintended lane departure. *SAE World Congress*, Detroit, MI, USA
- Qin SJ, Badgwell TA (2003) A survey of industrial model predictive control technology. *Control Eng Pract* 11(7):733–764
- Rajamani R (2006) *Vehicle dynamics and control*. Springer, New York
- Schorn M, Stahlin U, Khanafer A, Isermann R (2006) Nonlinear trajectory following control for automatic steering of a collision avoiding vehicle. In: *Proceedings of the American control conference*, 2006, 6 pp, 14–16 June 2006
- Shin M, Bae S, Lee J, Lee J, Heo S, Tak T (2003) New vehicle dynamics model for Yaw rate estimation. *Veh Syst Dyn* 37:96–106
- Sledge N, Marshek K (1997) Comparison of ideal vehicle lane-change trajectories. *SAE Trans* 106(6):2004–2027
- Soudbakhsh D, Eskandarian A (2010) A collision avoidance steering controller using linear quadratic regulator. In: *Proceedings of the SAE 2010, World congress & exhibition*, April 2010, Detroit, paper no: 2010-01-0459
- Soudbakhsh D, Eskandarian A (2011) Comparison of linear and non-linear controllers for active steering of vehicles in evasive manoeuvres. In: *Proceedings of the institution of mechanical engineers*, part I: *Journal of Systems and Control Engineering*, August 26, 2011, 0959651811414503, first published on August 26, 2011 (In press)
- Soudbakhsh D, Eskandarian A, Chichka D (2010) Vehicle evasive maneuver trajectory optimization using collocation technique. In: *Proceedings of the ASME dynamic systems and control (DSC) 2010 conference*, Cambridge, MA, 13–15 Sept 2010
- Talvala KLR, Kritayakirana K, Gerdes JC (2011) Pushing the limits: from lanekeeping to autonomous racing. *Annu Rev Control* 35(1):137–148
- Tideman M, van der Voort MC, van Arem B, Tillema F (2007) A review of lateral driver support systems. In: *Proceedings of the intelligent transportation systems conference*, ITSC 2007. IEEE, pp 992–999, 30 Sept–03 Oct 2007
- TRW (2011) Lateral support (LDW/LKA/LG). Available: http://www.trw.com/sub_system/lane_departure_warning_guidance_es. 31 July 2011
- Tseng HE, Ashrafi B, Madanu D, Allen Brown T, Recker D (1999) The development of vehicle stability control at Ford. *IEEE/ASME Trans Mechatron* 4(3):223–234
- Urmson C, Anhalt J, Bagnell D, Baker CR, Bittner R, Clark MN, Dolan JM, Duggins D, Galatali T, Geyer C, Gittleman M, Harbaugh S, Hebert M, Howard TM, Kolski S, Kelly A, Likhachev M, McNaughton M, Miller N, Peterson K, Pilnick B, Rajkumar R, Rybski PE, Salesky B, Seo Y, Singh S, Snider J, Stentz A, Whittaker W, Wolkowicki Z, Ziglar J, Bae H, Brown T, Demitrish D, Litkouhi B, Nickolaou J, Sadekar V, Zhang W, Struble J, Taylor M, Darms M, Ferguson D (2008) Autonomous driving in urban environments: Boss and the urban challenge. *J Field Robot* 25(8):425–466
- www.euroncap.com Ford lane keeping aid. Available: http://www.euroncap.com/rewards/ford_LaneKeepingAid.aspx. Accessed 31 July 2011
- Yih P, Gerdes JC (2005) Modification of vehicle handling characteristics via steer-by-wire. *IEEE Trans Control Syst Technol* 13(6):965–976
- Yoo H, Hunter D, Green P (1996) Automotive collision warning effectiveness: a simulator comparison of text vs. icons. *UMTRI*, Ann Arbor

Special Vehicular Systems

Ernst Pucher, Alfred Pruckner, Ralf Stroph and Peter Pfeffer

11 Drive-By-Wire

Alfred Pruckner¹ · Ralf Stroph¹ · Peter Pfeffer²

¹BMW Research and Development, Munich, Germany

²Hochschule München, Munich, Germany

1	Introduction	237
1.1	Vehicle–Driver Control Loop	239
1.2	Input Module Characteristics	240
2	Longitudinal Dynamic Systems	242
2.1	Functional Targets	243
2.1.1	Drive Functions	243
2.1.2	Brake Functions	244
2.1.3	Combined Function	244
2.2	Brake Pedal	246
2.2.1	Brake Pedal Characteristics	246
2.2.2	Decoupled Brake Pedal	247
2.3	Integrated Longitudinal Control	251
2.3.1	Autonomous Driving	252
2.4	Longitudinal Dynamic By-Wire Control Systems	252
2.4.1	Electronic Accelerator Pedal	253
2.4.2	Accelerator Force Feedback Pedal (AFFP)	254
2.4.3	Brake Force Feedback Pedal	254
2.4.4	ABS/ESP System	255
2.4.5	Electrohydraulic Combi Brake (EHCB)	255
2.4.6	Electrohydraulic Brake EHB, Simulator Brake Actuation SBA	256
2.4.7	Full Electromechanical Brake System (EMB)	258
3	Lateral Dynamic	259
3.1	Functional Targets	260
3.2	Steer-By-Wire Feedback Design	260
3.3	Lateral Dynamic By-Wire Control Systems	262
3.3.1	Vertical Force	263
3.3.2	Longitudinal Force	263
3.3.3	Lateral Force	265
4	Integrated Vehicle Dynamic	273
4.1	Corner Module	273
4.2	Control Strategy	274

5	<i>Functional Safety and Availability</i>	276
5.1	Basic Design of Safety Critical Systems	277
5.1.1	Safe State of a Vehicle with the System(s)	278
5.1.2	Danger and Risk Analysis	278
5.1.3	Degradation	279
5.1.4	Environment for Safe Drive-By-Wire	280
5.2	Example: Electronic Throttle	280
6	<i>Conclusion</i>	281

Abstract: Competitiveness to a company is given by innovations. The chassis as main part in vehicle design is incisive to the driving behavior of a car. On the one side, mechanical devices are well-engineered which means differentiation to competitors in mechanical devices is complex and costly. On the other side, improvements due to clients and legislator such as driving dynamics, CO₂ reduction, or pedestrian protection increase the requirements to the chassis concerning comfort, safety, handling, or individualization by less cost and maintenance.

This balancing act can be done by mechatronics systems which means the interaction of mechanic, electronic, and informatics devices. Basic mechatronics systems are used to assist the driver (e.g., power steering) or to overrule a wrong driver input (e.g., ABS brake). Different from this so-called by-wire systems are extensive mechatronics systems where the vehicle behavior and the driver feedback can be designed independently (there is no mechanical link between input and output).

Drive-by-wire, X-by-wire, or simply by-wire technology is already present nowadays. Starting with aeronautics, where fly-by-wire has been used extensively in the Airbus A320 family without mechanical backup. In passenger cars, by-wire functionality and by-wire systems are far more recent, but still already well known (VDI-Bericht 1828, 2004).

One can distinguish between by-wire functionality and by-wire system. The by-wire functionality can be reduced to the ability to control or even only apply a force by an electrical signal (through an electrical wire) to the vehicle. The definition by-wire system is that the line between the driver's input interface and the actuation which produces force is partly designed by wire. Hence, in contrast to the by-wire functionality, the system has no permanent hydraulic or mechanic linkage between them.

Common advantages of by-wire systems are the freedom in functionality, package integration, reducing variants, design, and enabler for driver assistance functions. In [▶ Sect. 1](#), these general facts of by-wire systems, the vehicle-driver control loop, and aspects of the input module behavior will be shown. Afterward longitudinal and lateral dynamic systems and their functionality are explained in more detail ([▶ Sects. 2 and ▶ 3](#)).

One benefit of by-wire systems is the system and functional integration. In [▶ Sect. 4](#) integrated corner modules are illustrated and analyzed as well as integrated control strategy aspects. The challenge of by-wire systems are the functional safety requirements, especially in terms of availability. The latter is the most important for the OEMs and customers. Any minor failure of the systems, which has to be displayed in the dash board, will reduce the customers' faith in the car. These aspects will be explained in [▶ Sect. 5](#).

1 Introduction

In vehicle technology, drive-by-wire means an electrical path between the driver input and the vehicle dynamics actuators. That means there is no mechanical link between driver input and tires, the vehicle dynamics are controlled by an electronic control unit. In order to reduce costs, mechanical fallback systems can be provided.

The big challenge on drive-by-wire is to get all the functional benefits compared to keep reliability and availability at acceptable costs. 📌 [Table 11.1](#) shows a summary of assets and drawbacks of by-wire systems.

One benefit of drive-by-wire is an easy realization of additional comfort and safety functionality. Comfort systems, such as (active) cruise control, park assistant, or variable steer ratio, will support the driver in normal driving situations. Situation-depending functionality such as variable steer ratio or optimal driving speed feedback on the drive pedal can be implemented easily. Safety functions such as antilock brake systems and dynamic control systems will work at incorrect use by the driver. Active safety systems like emergency braking, collision avoidance, or lane keeping systems will use by-wire functionality as well as future vehicle concepts with a high level of autonomous driving.

Another advantage is to achieve design space due to the omission of mechanical parts. This allows a tidy front package especially for the engine positioning, variation reduction due to communality of left and right steer systems, as well as improved passive safety conditions because of fewer parts which will impact the passenger area in case of collision. Common parts can be used in different concepts which enables a variation reduction, and differentiation can be done by software.

Due to the disconnection of input and output, the requirements to the Axle design can be reduced because wanted functions like directional stability and feedback to the driver can be done by software (Mueller 2010). By-wire systems allow suited use of energy which means a reduction of energy consumption.

Not least, by-wire enables totally new designs of input-systems like stick elements shown in 📌 [Fig. 11.1](#).

The challenge of by-wire systems is to provide high reliability with acceptable availability which means high redundancy requirements. Safety requirements in by-wire-systems without mechanical fallback means the system has to be fail-tolerant. In fail-tolerant systems, the driver has to be able to handle every possible failure which can occur (Winner et al. 2004). That means, all the electronic parts have to be configured at least twice or more often, which increases cost and might also result in more weight.

Unlike fail-tolerant systems, fail-safe systems include a mechanical fallback (Kilgenstein 2002). Different from electronic parts, mechanical parts can be designed in

📌 **Table 11.1**
Assets and drawbacks of by-wire systems

Assets	Drawbacks
Functionality	Cost
Enabler for active safety	Complexity
Design space	
Variation reduction	
Passive safety	
Simplified axes	



■ Fig. 11.1
Cockpit with stick element (BMW Group Research and Development)

a way that they will not break down when used as provided. However, an oversized steering column will not fail; an oversized electromotor will probably fail like a small one. An electronic system with mechanical fallback has to be designed as so-called fail-silent system. This means that in case of an electronic failure, the system will be shut down, the steering or braking function will be done restricted by the mechanical fallback system (Heitzer and Seewald 2004).

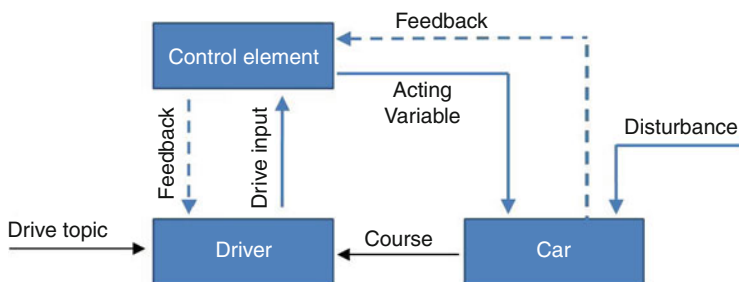
State-of-the-art antilock brake systems (ABS) are designed as fail-silent systems. The driver defines the deceleration of the vehicle by moving the brake pedal, which turns the pedal force into hydraulic pressure. If the wheel slip exceeds a certain value, the electronic control unit begins to reduce the hydraulic pressure independently from the pedal force. However, if ABS fails, the valves are set open; thus, a direct hydraulic linkage from the master brake cylinder to the brake caliper is given.

Superposition steering systems combined with power steering enable the control of steering wheel and steering torque nearly independent. Thus, most functional by-wire requirements can be achieved by these active steering systems, however, in most cases design advantages are not given due to the mechanical fallback.

1.1 Vehicle–Driver Control Loop

To learn more about the effect of by-wire systems, ► Fig. 11.2 shows the interaction of driver, vehicle, and environment in usual vehicles.

The driver controls the difference of drive topic and course by the control elements: steering, throttle, and brake pedal. The control elements transfer the driver input into the acting variables, longitudinal slip and slip angle, on the wheel. Haptic feedback of the control elements, mainly the steering feedback, gives information about the interaction vehicle – road and therefore information about the driving conditions to the driver.



■ Fig. 11.2
Driver, vehicle, and environment in usual vehicle

Beside the active variables, disturbances like side wind or changing road conditions act on the vehicle. The driver will be informed about the disturbances by the control elements as well.

In usual vehicles, mechanical systems such as steering wheel, steering column, steering gear, tie rod or brake pedal, brake cylinder, hydraulic brake caliper, and disk brake are defined as control elements. Electric, hydraulic, or pneumatic servo systems like power steering or brake booster will amplify the driver input. However, the functionality is warranted even if the support system will fail.

Different from usual systems, pure by-wire systems are characterized by a disconnected mechanical link between driver and acting variables to the vehicle. The control element is separated into input module and actuator module, as shown in

► Fig. 11.3.

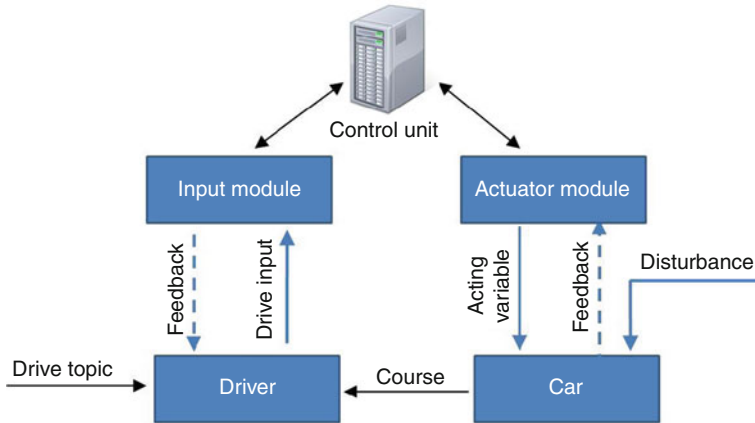
The input module has to detect the driver input by appropriate sensors and transmit feedback to the driver. The acting variables are adjusted by the actuator module which detects forces and displacements on the wheel. Both modules are communicating with the control unit where vehicle dynamic functions as well as safety requirements are controlled.

The input module characteristics, explained below, are significant for the quality of by-wire systems.

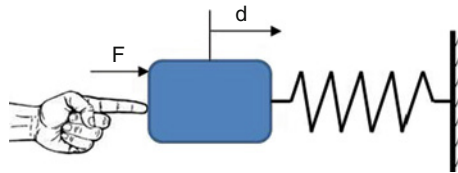
1.2 Input Module Characteristics

Passive input modules are simple and therefore cheap to realize. ► Figure 11.4 shows abstractly the context of input and feedback.

A proportional relationship between force F to the input module and displacement d is called isomorphic (spring centering) behavior and can be found, for instance, in usual brake pedals. The force on the pedal causes a pedal displacement, the actual force at the wheels, especially in slippery road conditions, is no return information on the pedal to the driver. That means even with locked wheels when the brake force is on the maximum limit the pedal force can be increased. Nevertheless, the proportion of force and displacement can be designed nonlinear in a way that in comfortable brake situations the force



■ Fig. 11.3
Driver, vehicle, and environment in by-wire-systems



■ Fig. 11.4
Passive input module behavior

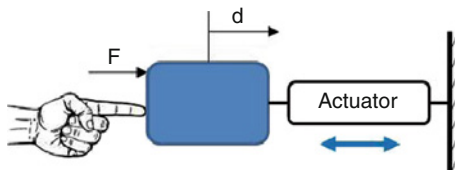
increment caused by a certain displacement input is small whereas in critical, high deceleration situations the force increment caused by the same displacement input is much higher.

If the spring rate in ● Fig. 11.4 is very small, the system is called isotonic (force-free) input module. The driver input is done by displacement, there is no force feedback to the driver. A nearly isotonic system is the gas pedal where the return force on the pedal is very small just to bring the pedal back to zero position. Usual drivers are not able to recognize force differences in the gas pedal, just the pedal displacement gives a direct haptic feedback.

On the other hand, isometric (displacement-free) input behavior means very high spring stiffness such as a touchpad. This isometric behavior is not practical for such a complex control task like driving a car.

Different from passive ones, *active input modules* are able to give feedback about the actual vehicle behavior to the driver. ● Figure 11.5 shows abstractly the context.

Concepts with force input and displacement feedback are using the measured force as input to the control unit whereas the resulting module position is placed by the input module actuator. According to this displacement, input and force feedback systems measure the displacement as control unit input and adjust the force and the input module.



■ Fig. 11.5

Active input module behavior

In usual linear vehicle dynamic behavior there is no difference between these two concepts; differences to the steering behavior in critical situations will be shown in chapter lateral dynamic (steer-by-wire feedback design).

An active input module behavior is given at state-of-the-art brake slip control systems. In general, the driver defines the wanted deceleration with the brake pedal position, the brake pedal force, and more than this the vehicle deceleration gives feedback to the driver as control input. Preventing locked wheels, the slip control system will reduce the hydraulic pressure in the brake system. This hydraulic modulation can be recognized by the driver on the brake pedal so that the driver becomes informed about the critical situation. However, the driver must not lift the pedal due to this force modulation to achieve the best deceleration possible.

Active throttle pedals can be adjusted depending on the actual driving situation. By means of an active throttle pedal with variable kickdown position, the actual speed limit information or the optimal difference to the vehicle in front can be given to the driver haptically. Of course, the driver has to be able to overrule this recommendation.

Since one's hand is more sensible than the feet, velocity control per stick can be handled different from pedal input. The required acceleration or deceleration for instance is measured by the force on the stick, the actual vehicle speed is given back to the driver by the stick angle. Active steering systems may define the wanted lateral acceleration by the steering-torque input, the resulting yaw rate information is given back to the driver as steering wheel angle. More about brake and steering functionality is explained in the next two chapters.

2 Longitudinal Dynamic Systems

Longitudinal vehicle dynamic systems can be divided into drive and brake systems. The main input interfaces are drive and brake pedals; secondary interfaces are gear shift systems, which are not addressed here.

One can distinguish between by-wire functionality and by-wire system. The brake-by-wire functionality can be reduced to the ability to control or even only apply a longitudinal dynamic force by an electrical signal (through an electrical wire) to the vehicle.

The definition brake-by-wire system is that the line between the drivers brake interface and the actuation which produces brake force is partly designed by wire. Hence, in contrast to the by-wire functionality, the system has no permanent hydraulic or mechanical linkage between them.

2.1 Functional Targets

► **Table 11.2** shows an overview of drive and brake functions and their quality in a passenger car:

2.1.1 Drive Functions

Drive-by-wire functionality has first been introduced in passenger cars as cruise control with mechanical linkage. Hence, in some cars, the throttle pedal moved as the cruise control changed the amount of combustion engine torque. Other solutions were realized

■ **Table 11.2**
Significant longitudinal dynamic functions

		Function	Quality
Drive functions		Traction control	Wheel spin prevention
		Cruise control	Constant vehicle velocity control
		Optimal speed information	Best speed according to law, traffic and vehicle dynamics
Brake functions	Safety functions	Antilock brake control	Best stopping distance at any road condition (μ -high, μ -low, μ -split) plus steerability and stability
		Dynamic drive control	Vehicle dynamics control by individual brake intervention
		Brake assistant	Faster and higher brake forces in case of fast pedal use
		Acceleration control	Stability and optimal acceleration
	Assistant functions	Active cruise control	Active distance control
		Hill assist	Starting aid on hill situation
		Emergency brake	Driver independently system brake intervention
Combined function		Energy recuperation	Optimal energy recuperation at maximum vehicle stability in hybrid and electro vehicles

with an electric motor at the throttle flap which kept the pedal at the idle position while the cruise control was adjusting the combustion engine torque to keep the desired speed.


The next step was using the freedom of the feedback in the driver's throttle. Optimal speed information could give active suggestions of the best throttle pedal angle. This could be used for economic driving, for speed limit assistance according to the law, or even according to the navigation information on the right speed of the next curve.

2.1.2 Brake Functions


Today's brake systems go farther than reducing the velocity of a vehicle according to the driver's input via brake pedal. In modern passenger cars, it is expected that the brake system is also able to cope with different street situation, such as high- μ or low- μ and any changing transition in between, especially μ -split (antilock brake systems). Furthermore, braking in curves and close to the critical area of high driving dynamics is controlled and supports the driver in his driving task (dynamic stability control). Brake assistants reduce the brake distance by amplifying the driver input in detected emergency situations.

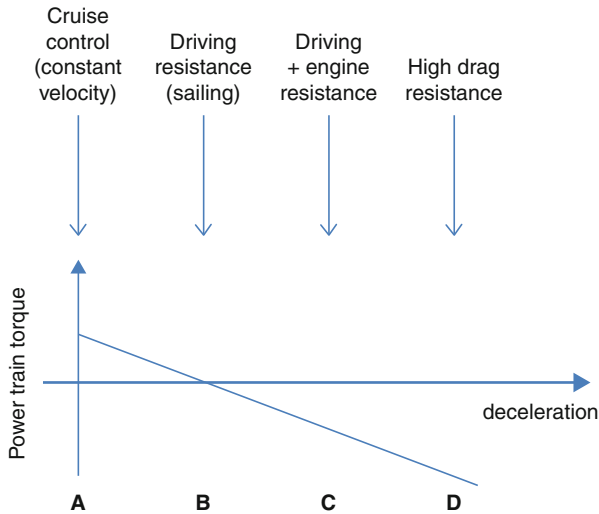
Even though the brake system is not intended to be used for acceleration, it can help the car accelerating with wheel-selective brake force on slippery surface, especially on μ -split, where one driven wheel would be on low- μ and the other on high- μ . With a differential, the low- μ wheel would just slip and the car would not accelerate, as it does when the low- μ wheel is applied with a brake force and the torque can actually be divided and be used on the high- μ side to accelerate the car.

The other area of supportive systems are the driver assistance systems, such as cruise control, distance control, hill hold or hill assist, which take the advantage of the ability of the modern brake system to provide a brake force "by wire" from an external source. Also the brake system as it is designed already is able to make an emergency brake independent of the driver, if other systems are detecting a situation which might force this action to limit damage.

Combined drive and brake functions are needed in powerful hybrid electric vehicles (HEV) and battery electric vehicles (BEV), respectively. The power of the electric motor acts as an additional degree of freedom in HEVs for the MMI. Also considering BEVs there are new possibilities to design the MMI. In  Fig. 11.6, the different possibilities for the design of the BEV MMI are plotted and explained hereafter.

2.1.3 Combined Function

In situations the driver neither actuates the throttle nor the brake pedal, a certain torque acts on the power train. The possibilities vary between a positive torque that acts on the power train in a way that the current vehicle speed is kept constant (point A in  Fig. 11.6), no torque input so that the vehicle decelerates very slow due to the driving resistance (B) and the overrun fuel cutoff, in which a certain negative torque acts on



■ Fig. 11.6

Different recuperation strategies while no pedal contact, HEV and BEV

vehicle and power train and decelerates the vehicle. In HEV's and BEV's, this overrun is realized by regenerative torque of the electric motor in order to store the vehicle's kinetic energy in the battery. The absolute value in that condition needs to be defined. It varies between values similar to conventional cars (C) and a high amount, which allows decelerations that suffice common driving situations (D). The value is limited by the technical parameters of the electric motor and electrical system as well as the vehicle dynamics (Eberl et al. 2011).

Considering the fuel consumption, regenerative braking needs to be maximized in all drive situations. Systems with recuperation strategy D (high drag resistance) decelerates up to 2 m/s without using the brake pedal. This allows recuperating most of the kinetic energy in usual driving situations; the brake pedal will be used rarely, mainly in emergency situations. Therefore, the brake pedal can be coupled to the usual brake system which means less costs. Due to the fact that the driver controls drive and usual brake situations just by the drive pedal, the drive feeling is very different from nowadays vehicles. When using just one axle to decelerate the vehicle, the driving stability is worse especially in case of rear drive (oversteer tendency). At last, the deceleration depends on the battery state; if the battery is cold, no recuperation can be done and therefore the vehicle behavior becomes varying. To prevent varying brake characteristics and to be able to use different recuperation strategies, a decoupled brake system with actively varying brake pedal feel and full brake-by-wire functionality is needed.

Additionally, in critical driving situations such as braking on low- μ and hence using the ABS or braking while the brake system is stabilizing the vehicle, the driver usually also uses the brake pedal. Therefore, two commands are acting on the same hydraulic brake actuator, which give implications in both functions. The control function should not

ignore the driver input and therefore his/her desire to control the vehicle and the driver still wants the help of the brake system, such as the ABS control. Therefore, a decoupled brake system is not essential but helpful to design good brake functions with acceptable feedback.


2.2 Brake Pedal

More than the drive pedal, the brake pedal is the important longitudinal vehicle HMI. Brake pedal characteristic is not only described by the pedal force with results from a certain pedal travel, but also from the deceleration the driver feels when applying the brake pedal. Therefore, the principal brake pedal characteristic is explained in detail followed by the functional description of using decoupled pedals.

2.2.1 Brake Pedal Characteristics

The brake pedal feel can be described by several characteristics, which will be briefly explained (Breuer and Bill 2003).


Pedal Force Versus Pedal Travel

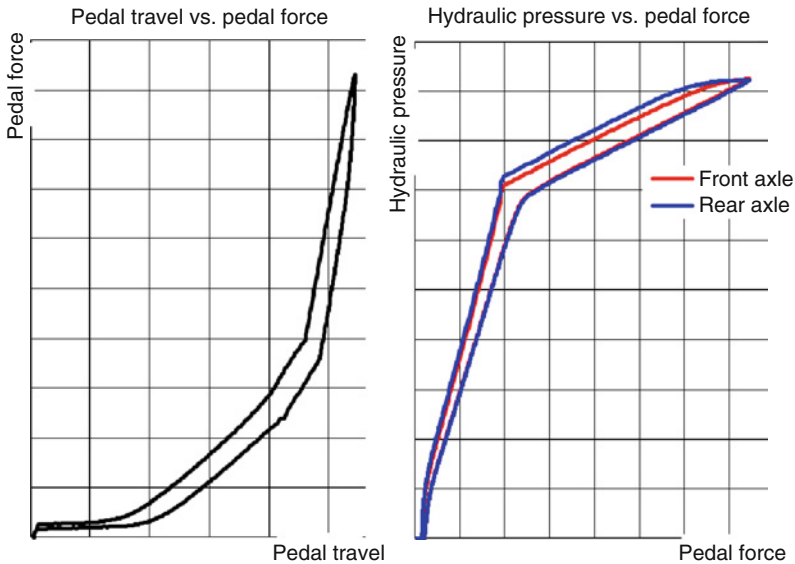
In order to apply the brake pedal in a sensitive manner, the correlation between the pedal force and the pedal travel should have a small increase in the force, while having a relatively large travel during low deceleration and exact dosing of the brake force. For higher decelerations up to an emergency brake, higher forces should be used, as the deceleration also puts an additional force on the extremities which apply the pedal force. Compare the left-hand side of  Fig. 11.7. The total pedal travel is normally limited by the assigned package for the pedal box.

Damping and Hysteresis

When controlling the deceleration, damping and hysteresis is useful for smoothing the input variations while disturbances, like potholes or vibrations, take effect from the environment. However, in the case of very fast appliances, this is an unwanted behavior.

Amplification Factor of the Brake Booster

The brake booster is normally designed, that it supports a high- μ maximum deceleration with a certain pedal force, every person can reach easily. In most passenger cars, it is around 150–250 N at the brake pedal for a deceleration of 10 m/s, rather than just reaching the criteria of the homologation as mentioned above. As shown on the right-hand side of  Fig. 11.7, the amplification is the gradient of the pressure over the pedal force, which goes down to no amplification when the vacuum of the brake booster is exhausted.



■ Fig. 11.7

Pedal force over pedal travel and hydraulic pressure over pedal force

Jump In

In conventional vacuum brake booster, the jump in means a quick increase in brake liquid pressure when pressing the brake pedal over a certain point. This gives the driver a feel of a good response of the brake system. In very low speed and braking force situations, it can be annoying.

Statute Law

The brake pedal characteristic of a conventional hydraulic brake system with a pneumatic brake booster is well established. Some of the characteristics that are in such systems are only a result of the system and its limitations, such as damping, stiffness, volume consumption, level of vacuum, etc. Others can be influenced by choosing the size of the components, their mechanical geometries, and the design of their springs. However, one main design criteria for the brake system is the law for homologation ([ECE R 13](#)), where a deceleration of 0.63 g with a pedal force of 150 N and in case of a failure of the brake booster a deceleration of 0.32 with a force of 500 N at the pedal.

2.2.2 Decoupled Brake Pedal

Having learned the main characteristics of a conventional vacuum brake booster system, one can decide which of the mentioned useful and parasitic properties should be rebuilt in a brake-by-wire system where most of the characteristics can be chosen individually and

with a high degree of freedom. Certainly, trying to match the exact pedal feel of a conventional brake system will have the least trouble to get used to the new system, and therefore its acceptance in the market will be high.

The difference between brake-by-wire with an energetically coupled fallback level or without is not so much the functionality itself, but the degree of freedom one has in designing the pedal feel and also the limitations which are present when a fallback level has been designed. Those limitations are often a much higher pedal force or pedal travel or a combination of both, which clearly must achieve the regulations, but can often hardly be handled by a normal driver who is surprised by the change.

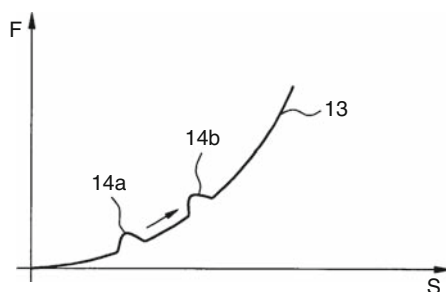
It might also be useful to adapt the pedal feel to different driving situations. Nowadays, the driver does get a pulsation of the brake pedal in case of ABS activities. Also, one could wish for a different pedal feel characteristic depending on its speed, the mass of the vehicle, or just his/her preference as being more sporty or rather comfortable driver.

Different additional functions could need a different pedal feel. Maybe drivers of BEV with high regenerative thrust want a different feedback than in conventional cars. They might want to know when they are using regenerative braking or braking with a hydraulic/electromechanic brake. ● [Figure 11.8](#) shows a discontinuity in the pedal force over the pedal travel, similar to a kickdown, known in acceleration pedals of automatic drive cars (Prickarz and Bildstein 2009). It could be used to show the driver an optimal deceleration for regenerative braking, which might change due to velocity, battery charge, and other inputs to the regeneration strategy. Hence, the discontinuity can move, like from point 14a to 14b.

Brake systems with decoupled brake pedal give the freedom of implementing some functions which are not possible without any pedal implications in conventional systems. Next, these functions are explained in detail.


Regenerative Braking

In powerful hybrid electrical vehicles or in electric vehicles powered by battery, fuel cell or similar, regenerative braking is a very powerful function to change the kinematic energy of the moving vehicle into electric energy and store this energy again for a following acceleration.



■ Fig. 11.8

Pedal force over pedal travel for regeneration feedback (Prickarz and Bildstein 2009)

Regenerating energy is normally done by the means of using the electric motor as an alternator which is driven by the rotation of the wheels. The power electronics controls then the power and the torque of the electric motor to decelerate the vehicle. The main issue is now the control of the deceleration according to the driver's input. As described above, the behavior of an electric vehicle could range from a one-pedal handling with high deceleration forces when lifting off the drivers pedal, via the well known and used to behavior with a decelerating force of about the same amount of a conventional vehicle in gear, up to zero deceleration, hence maintaining the velocity when neither pedal is actuated and therefore only decelerating when the brake pedal is pressed (compare  Fig. 11.6).

The control strategy of the electrical power management and the driver feedback and HMI is the main factor when choosing the strategy for regeneration and the functions concerning the regeneration.

For the first strategy of using only one pedal for acceleration and deceleration with the electric motor, the brake system and brake functions do not have to be changed. Problems might arise when regenerating on low μ followed by increased wheel slip up to wheel lock situations due to the high inertia of the electric motor. Then, the ABS function is not anymore that straightforward, to lower the brake forces and get the wheel spinning again, as the high inertia keeps the wheel slip at a high level for a long time. Algorithms are then sought to control the electric drives, by either decoupling the electric motor, if possible, or contributing an accelerating force to the drive train which works against the inertia, or in the case of electric motors directly at the wheels, to implement a slip controller in the electric drive controller (Teitzer et al. 2010).

To keep the strategy close to today's known deceleration due to the normal engine idle torque, the regeneration could either be hidden within the first couple of millimeter of the brake pedal travel, or with higher electrical power, the brake pedal needs to be decoupled from the brake system in order to hide the pedal implication or the deceleration implications when blending from electric to hydraulic deceleration. Such decoupled systems are explained later.

Cruise Control with Braking

Another function which can be improved, but is not driven by electric drives or hybrid vehicle architectures, is the driver assistance which makes use of the brake system. One very prominent system is the cruise control with braking function or in a higher variation the adaptive cruise control which also controls the distance to the previous car.

When such a function is implemented without a brake-by-wire functionality with a decoupled brake pedal, the brake pedal characteristics change in general during the actuation of the brake system. The most common realization is by applying a hydraulic pressure using the pump from the ESP system which applies the hydraulic pressure. In the case of an overtaking the control by the driver pressing the brake pedal, the pedal feel characteristic is different as hydraulic pressure is already in the system and the brake pedal force in relation to the pedal travel is therefore higher when using the brake pedal to increase the brake force any further.

Using brake-by-wire with a decoupled brake pedal, the pedal feel characteristics are largely independent from the applied brake force at the brake caliper and therefore have no implications whether the cruise control is actually using the brake system or not. The only implication could be the strategy, as to when the cruise control function is deactivated. For example, it could be either when the driver applies a small force on the brake pedal, which could result in less deceleration than before, even though the brake pedal has been applied, or when the deceleration of the driver input would result in a higher deceleration force than it is applied by the cruise control system.

Returning to the main goal of the cruise control systems, maintaining not only speed, but also reducing speed when ordered by the driver via a hand-controlled switch or lever, or when the previous vehicle reduces its speed or changes with a lower speed into one lane. Driver implication via the brake pedal is not the only advantage of a by-wire brake system. The very fast and exact by-wire brake actuators that control the brake forces can make very smooth decelerations and also control very exact a speed due to a small error. This is very important when driving with cruise control downhill and the speed can only be maintained by permanent brake apply.


Antilock Braking/Dynamic Stability Control

One last example, even if definitely not a major driver for a decoupled by-wire brake system is the stability control as well as the ABS function. Also here, the two aspects of the driver implications and the control performance and smoothness of the brake applies can be said.

Starting this time with the performance, by-wire brake systems are in general faster in applying the necessary brake force and therefore are quicker in stabilizing the vehicle. Also the ABS with by-wire system, which, in general, can control the brake force and therefore the wheel slip almost analog and hence in very fine steps, is of a higher performance and might reduce the stopping distance.

Whether a driver wants a feedback in the event of an ABS braking situation via brake pedal is a psychological question which will not be tackled here. It is clear, with a by-wire system, stabilizing brake applies cannot be felt directly in the brake pedal, as it is the case in today's conventional brake systems. Unless the by-wire system is extended by an active pedal simulator, which can not only reproduce a given brake pedal characteristic, but also can change the brake pedal force actively, or in a less sophisticated manner, just vibrates in the case of a stabilizing brake action.

Different HMI for Brake Commands

The above-mentioned functions are all assuming that the driver controls the brake with a brake pedal. By-wire without a backup path can also use any form of input, which can be changed into electrical commands. This has already been used to build prototypes with a side stick control (compare  Fig. 11.1). Also for special vehicle, for example, for people who cannot use their legs to control the brake, it could be controlled by a joystick, hand pedals at the steering wheel or even through speech commands, even though it is not feasible to our knowledge.

2.3 Integrated Longitudinal Control

The next use of by-wire functionality in longitudinal control is the combination of positive and negative acceleration into one integrated longitudinal control of the vehicle. Following the use of the integrated by-wire functionality is described from today's well-known eGas with a conventional brake system up to a combination of several by-wire systems.

- *Traction slip control system:* In today's modern vehicles, the integration of the combustion engine control and the brake system control is already implemented in the functionality of the accelerating slip control, especially during μ -split acceleration. When a driver demands more positive wheel torque than the wheel is able to transfer to the road surface, the wheel would slip and the vehicle would become instable and not accelerate at its maximum possible rate. First, when detecting a too high wheel slip at the driving axle, the control system demands to reduce the engine torque in order to reduce the wheel slip, without any action of the brake actuators. If this is not enough, or not the best control strategy, the brake system supports the controller by applying a brake force and therefore reducing the wheel slip. In case of μ -slip acceleration, the wheel on low μ would spin and the wheel on high μ would not get enough torque to accelerate the vehicle, due to the open differential gear. Hence applying a brake force to the spinning wheel, the engine torque would accelerate the vehicle with the wheel on the high μ surface.
- *Traction slip control system for HEV:* Having a vehicle, which has an additional electric motor within its power train, as most hybrid electric vehicles do, the above-described controller has now a third actuator to control the acceleration, which is commanded with the one-throttle pedal. The controller must decide whether to reduce the combustion engine torque, the electric motor torque, or use the brake force to control the wheel slip in order to accelerate and keep the vehicle in a stable situation.
- *Antilock and regenerative braking for HEV:* Using the above example of vehicle, it is usually also capable of applying a decelerating force with the electric motor by regeneration, hence using the motor as an alternator and regenerating the energy into electric energy, which can be stored. Therefore, the vehicle does not only decelerate due to the force of the combustion engine and the brake force of the brake system when the brake pedal is applied, but also by the regenerative torque of the electric motor acting as an alternator. In powerful hybrids or electrical vehicles, it is advisable to have a decoupled brake system, in order to control the decelerating force of the brake system and the electric motor in a manner which the driver does not recognize. Also, depending on the power of the electric motor, the regenerative torque can already drive the wheels into wheel slip, which must be reduced by reducing the decelerating torque of the electric motor. Hence, the motor must perform an ABS control strategy, which of course must be able to interact with the ABS controller of the brake system. For such a complex system, an integrated longitudinal controller to control all four actuators would be best.

- *Antilock and regenerative braking for BEV*: For pure battery electric vehicles, the above is valid, except, that no combustion engine needs to be controlled. However, the electric motor is normally much more powerful and especially the control of the deceleration must take the electric power into consideration for an optimal control of the vehicle.

2.3.1 Autonomous Driving


All the above functions and techniques are the base for autonomous driving. The “drive-by-wire” functionality itself is sufficient for autonomous longitudinal driving. Each system has constraints and the degree of by-wire functionality depends on the application and therefore on the actuation speed limits and reliability. Also the drivers’ implications would be huge if all systems were realized with mechanical backup. Moreover, if the driver is no longer in the loop for driving, the functional safety requirements of the by-wire systems have to be almost the same as without mechanical backup, as the driver cannot react quickly enough when autonomous driving is performed (Eckstein 2001). For example, a helicopter pilot, when flying with autopilot, is allowed a 3 s reaction time, after a fault occurred, before taking the controls. After that, the helicopter must be controllable by the pilot with a moderate amount of pilots’ workload. For a passenger vehicle, with other vehicles very close in heavy traffic, a lateral displacement or a delayed reaction time for braking could be fatal.

2.4 Longitudinal Dynamic By-Wire Control Systems


In the previous chapters, the basic and more sophisticated longitudinal by-wire functions were explained.

Now, the systems which are able to realize those functions will be explained in a little more detail. Starting again with well-known systems and ending with a full brake-by-wire system.

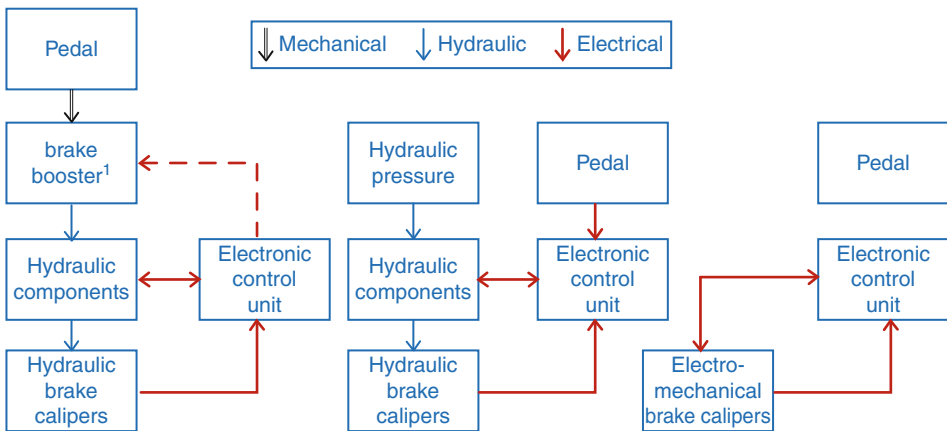
Brake-by-wire means that the connection between the brake pedal and the brake actuation is partly or completely realized by an electrical wire.

For a quick overview of the systems which can be called as “by-wire,” a summary of them is presented here. The definition brake-by-wire system which has been given in the introduction is that the connection between the driver’s brake interface and the actuation which produces brake force is partly designed by wire. Hence, in contrast to the by-wire functionality, the system has no permanent hydraulic or mechanic linkage between them. In  Fig. 11.9, the different systems are shown with their means of energy medium, which is used for their function. Hence, there is no more mechanical or hydraulic linkage between the driver’s foot and the brake actuators (brake calipers in conventional cars). This also results in the fact that the brake pedal feel has to be done by other means than the usual hydraulic feedback as already mentioned previously.

2.4.1 Electronic Accelerator Pedal

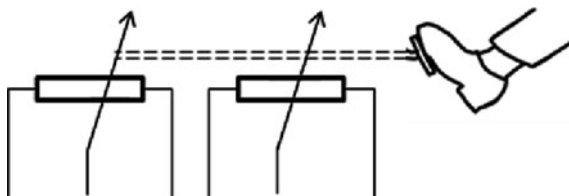
Drive-by-wire in the means of the electronic throttle without a mechanical linkage is nowadays very common and well established.  *Figure 11.10* shows the principle drivers input recording by two redundant potentiometers.

One of the advantages is the freedom of mechanical integration as well as the freedom in changing the motor torque independent of the drivers input. Also, the actuation does not need to move the whole mechanical linkage, but only control the combustion engine torque, which is, in today's modern cars, done by the amount of fuel injection and is therefore much quicker. Moreover, the damping of the accelerator retraction, which is also pollutant-reducing, is now more easily possible. Even the reduction of engine torque with fully pressed accelerator is possible, for example, when the automatic mode shifts during an acceleration process. Of course, there are also other possibilities for influencing the engine torque, for example, the very fast reacting ignition adjustment and the slower charging-pressure adjustment.



¹pneumatic, hydraulic or electric

■ **Fig. 11.9**
Schematic overview of different brake-by-wire systems



■ **Fig. 11.10**
Drive-by-wire driver's request recording with two redundant potentiometers

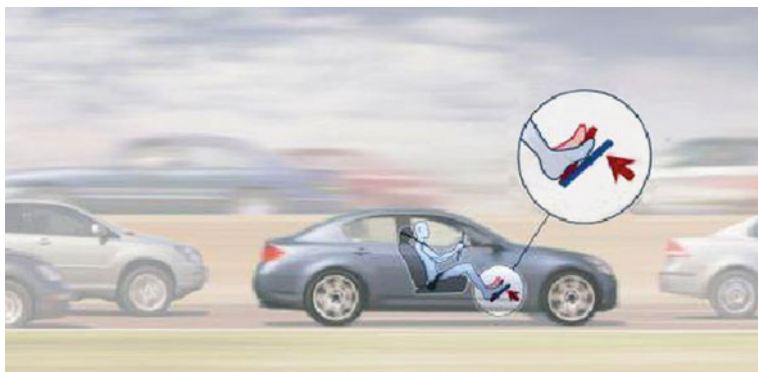
2.4.2 Accelerator Force Feedback Pedal (AFFP)

The next step was using the freedom of the feedback in the driver's throttle. An active force feedback pedal could give active suggestions of the best throttle pedal angle with the help of opposing force in the accelerator pedal. This could be used for economic driving, for speed limit assistance according to the law, or even according to the navigation information on the right speed of the next curve. Recently released by Continental is the accelerator force feedback pedal (AFFP) that gives warning of dangerous situations by vibrating and exerting counterpressure in the accelerator pedal (🔗 [Fig. 11.11](#)). This should make the driver take his foot off the pedal and get ready to brake.

The pedal can also help to drive at a more even speed and therefore more economically with the aim of reducing fuel consumption and CO₂ emissions. The system uses information from the radar or camera sensors, identifies the best speed for staying with the flow of the traffic, and warns the driver by gentle pedal counterpressure that if they were to accelerate any more, the driver would be exceeding the optimum speed range and distance. The result is that the vehicle keeps to an even engine speed, avoiding frequent braking and accelerating. However, the driver still remains in control and can accelerate if required. The AFFP also offers advantages for hybrid and electric vehicles. In a hybrid, it could give feedback to the driver warning that an approaching pedal position would activate the internal-combustion engine. In electric vehicles, it could indicate how the driver's vehicle operating profile was adversely affecting battery range.

2.4.3 Brake Force Feedback Pedal

Another freedom in brake pedal feel is the active force feedback brake pedal. Thus far, this kind of interaction with the driver has only been used for the throttle (as described above) or for experimental cars in which the brake pedal feel has been adapted actively to examine



■ Fig. 11.11

Accelerator force feedback pedal (AFFP) (Zell et al. [2010](#))

the subjective feeling of different pedal feel characteristics for the driver. However, as in [◆ Sect. 2.2.2](#), some function have nowadays an implication on the brake pedal, which is not actively controlled as one would desire, but as a parasitic result of the limitation of systems with a mechanical or hydraulic linkage. One prominent example is the pulsation of the brake pedal when being in an ABS situation, which gives the driver a direct force feedback of the road condition.

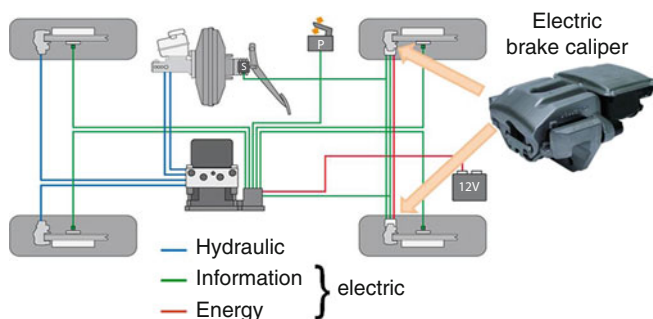
2.4.4 ABS/ESP System

Starting with the best known and most common system for braking, the antilock function (ABS) and its extension to the electronic stability program (ESP) or dynamic stability control (DSC), as it is also called. The ABS and ESP are not classic brake-by-wire systems, as the driver does not use these systems directly for braking. However, it has most of the functionality of a brake-by-wire system, as it can, independent of the conventional hydraulic linkage between the brake pedal and the brake calipers, change the deceleration of the vehicle, for the ESP in both directions. As these alterations are done electronically, we could call the system therefore “by wire.” One little step further is the cruise control with active braking. It is often realized with the ESP or DSC and hence it is a brake-by-wire system. The driver can reduce the speed, which might be done by applying brake forces at the vehicles wheel brakes, by using a button or lever to adjust the desired speed. Therefore, the connection between the drivers input of deceleration, here the desire of reducing the speed, and the brake actuator which applies the brake force is by wire. However, as the brake pressure is also altered at the main brake cylinder, the driver does have pedal implications when using the brake pedal in such a situation where the system applies a brake force. Therefore, one does not call the ESP a by-wire system.

2.4.5 Electrohydraulic Combi Brake (EHCB)

The electrohydraulic combi brake, developed by Continental automotive, is a partly by wire and partly conventional brake system. With its combination of a hydraulic brake at the front axle and an electromechanical brake at the rear axle, proven and safe conventional brake system technology is combined with decoupled brake-by-wire functions. [◆ Figure 11.12](#) shows the concept of an electrohydraulic combi brake system with hydraulic brakes at the front and electric brake at the rear side, taken from the Continental Media Center. Additionally, an example of an electric brake caliper is shown. Depending on the vehicle mass and the position (front or rear vehicle side), the system, in general, can be used with 12 or 42 V. Due to the integrated electromotor, the caliper is bigger than a hydraulic caliper.

With such a system, the previously described functions without pedal implications ([◆ Sect. 2.2.2](#)) can be performed on the axle with the electromechanical brake actuators. Hence, using this axle for regeneration and brake blending or for external brake demands such



■ Fig. 11.12

Concept with front hydraulic and rear electric system (Continental Media Center)

as the cruise control or even for standstill management, no direct brake pedal implications will be apparent for the driver. However, if the brake force distribution must be changed from the initial 0% to 100% distribution, which is the requirement for elimination of pedal implication, the hydraulic brake on the other axle needs to be applied, resulting in an increasing pedal force of the brake pedal. Reasons for changing the brake distribution are wheel slip, higher brake forces, steering inputs, etc.

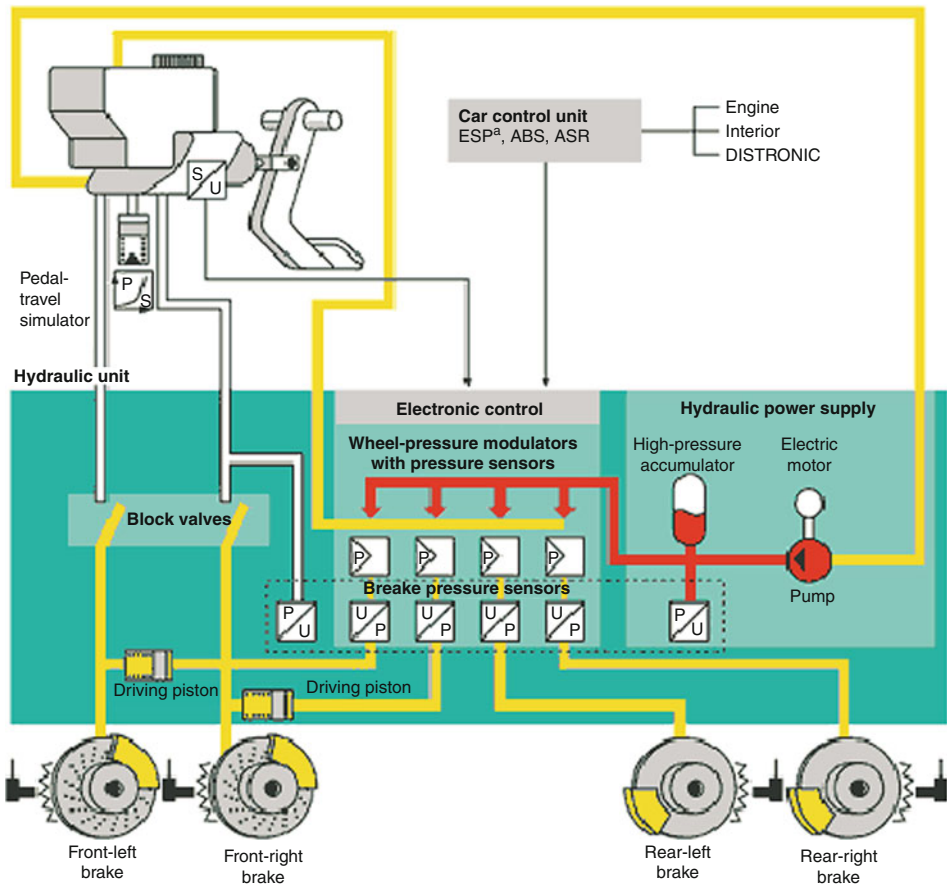
2.4.6 Electrohydraulic Brake EHB, Simulator Brake Actuation SBA

Having learned the disadvantages of controlling only a part of the brake system by wire, the next increment is a system, which has the by-wire functionality without pedal implication on all four wheels, and does still have the hydraulic fallback, that allows the driver to achieve a decent brake force when the electronic of the brake system has failed. Two such systems are currently in series production vehicles, especially for hybrid electrical vehicles. One is the better known electrohydraulic brake system (EHB), for example from Bosch (also called Sensotronic Brake Control (SBC) by Daimler) or as Electronically Controlled Brake ECB from Advic. The other known system is the Simulator Brake Actuation (SBA) from Continental.

➤ Figure 11.13 shows the principle of the sensotronic brake control. The system provides the brakes with a fluid supply from a hydraulic high-pressure reservoir sufficient for several braking events. A piston pump driven by an electric motor enables a controlled brake fluid pressure of 14–16 MPa (2,030–2,320 psi) from a gas diaphragm reservoir. The brake pedal movement is detected by a sensor and the brake pedal feel is reflected by a simulator. When the brakes are activated, the EHB control unit calculates the desired target brake pressures at the individual wheels. Braking pressure for each of the four wheels is regulated individually via a wheel pressure modulator, which consists of one inlet and one outlet valve controlled electronically. Normally, the brake master cylinder is detached from the brake circuit, with a pedal travel simulator creating normal pedal feedback.

Sensotronic Brake Control: The high-pressure brake system of the new Mercedes-Benz SL-Class

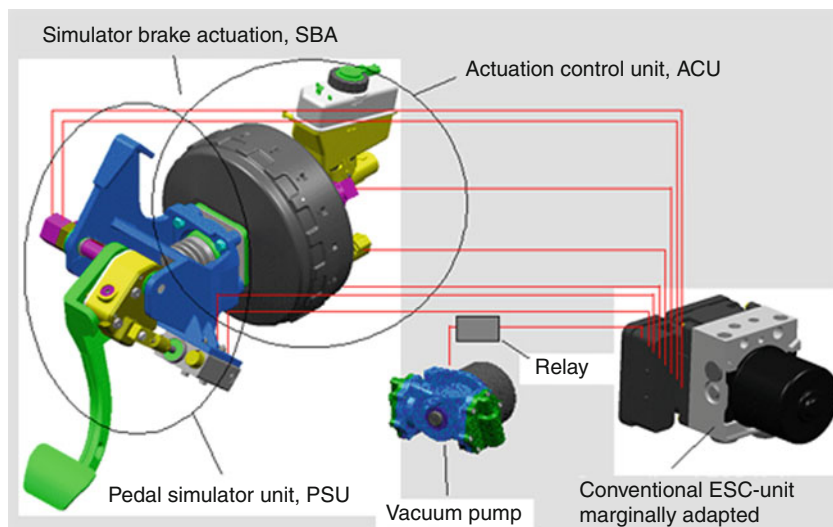
Actuation unit



■ Fig. 11.13

Sensotronic brake control (Fischle and Heinrichs 2002)

For the Simulator Brake Actuation SBA (► Fig. 11.14), the hydraulic control unit and the vacuum booster are almost similar to a conventional system. The booster is a further developed active vacuum booster, as it is used to apply brake pressure when using driver assistance systems such as adaptive cruise control. The main part is a gap in the mechanical linkage between the brake pedal and the input to the vacuum booster. With no mechanical linkage, the brake pedal feel has to be simulated, which is done by a dry simulator with spring and an elastomeric block within the pedal simulator unit. When the brake pedal is pressed, a pedal sensor sends the deceleration demand to the electronic control unit (ECU). This unit controls the active booster to fill the hydraulic brake system with the appropriate pressure, altered by any additional function, for example reducing



■ Fig. 11.14
The SBA (simulator brake actuation)

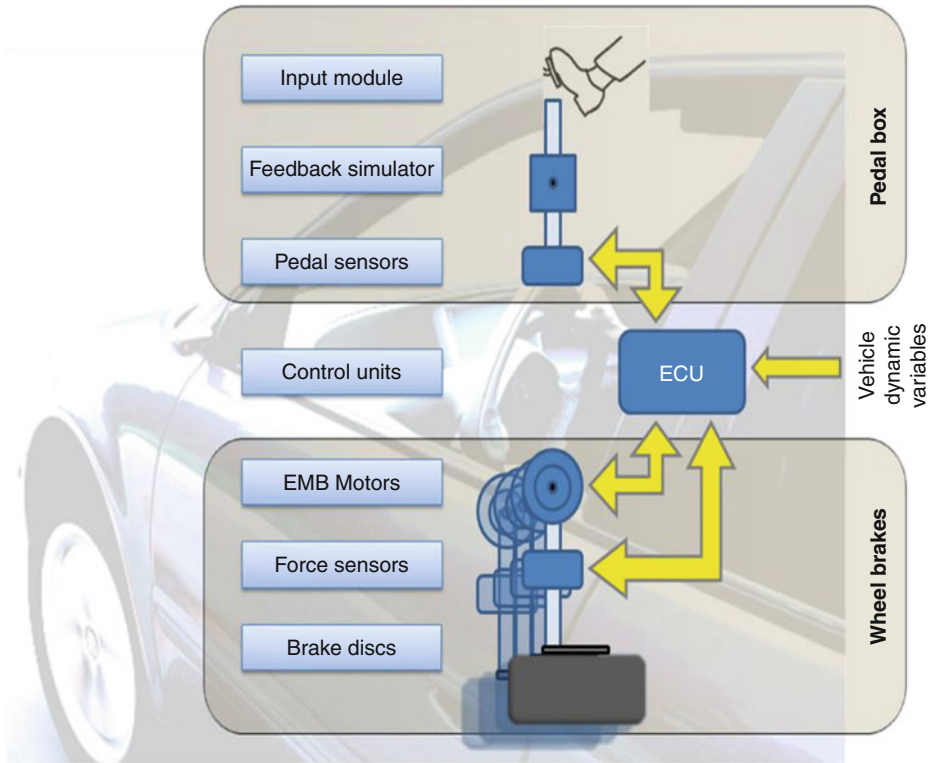
the brake pressure for blending with the regenerative braking without any implication at the pedal, as described above. The restriction of the amount of blending is designed into the system by the size of the gap in the mechanical linkage.

When a system failure is detected, the pedal feel simulator is switched off and the driver can depress the brake pedal to overcome the gap in the mechanical linkage. This applies the force to the vacuum booster and therefore the hydraulic brake system with a similar pedal feel but longer pedal idle travel is given. Figure and details are taken from [Conti homepage – SBA](#).

2.4.7 Full Electromechanical Brake System (EMB)

The ultimate brake-by-wire system is a system to apply brake forces directly at the brake discs without any hydraulic liquid and by any means of input. Such a system has no mechanical fallback. Most commonly, the brake-by-wire system consists of four electromechanical brake actuators (see ► [Fig. 11.12](#)), one at each wheel. Directly at the actuators electronic control units are placed and connected by a bus system to the central control unit.

The driver's input might be detected by a side stick or for autonomous driving by sensors which detect the environmental hazards and the programmed speed and therefore deceleration. However, a very convenient start for brake-by-wire vehicles will be a dry brake pedal simulator unit which senses the driver's desired deceleration. Such a system has all the freedom to implement the above-mentioned functions without any restriction in their amount without any pedal (or any other HMI) implication (► [Fig. 11.15](#)).



■ Fig. 11.15
Full electromechanical brake system

3 Lateral Dynamic

The steering input on the front axle is the main input variable to the lateral dynamics. The driver controls guidance and stabilization by the lateral tire force and gets information about the road surface by the steering system.

The task for the driver can be defined as:

- Definition of the vehicle course
- Low-frequency course correction due to disturbances
- High-frequency vehicle stabilization due to changing road conditions or wrong driver input

Mechatronic systems enable improvements in the vehicle design to support the driver in different driving situations, for example, variable steering torque and steering ratio may improve the comfort to the driver. Lateral dynamic interventions reduce the accident statistics. Road condition feedback can be given selective by mechatronic systems. In future, mechatronic systems can be used for active crash avoidance up to autonomous driving.

3.1 Functional Targets

By means of the tasks of mechatronical systems, functional targets can be defined. The following table shows the most significant functions of controllable steering systems and their qualities (➤ [Table 11.3](#)).

Beside the functional goals in ➤ [Table 11.2](#), lateral dynamic systems have to be safe and available as well as profitable, efficient, flexible, quietly and diagnostic able (Fleck et al. 2001).

3.2 Steer-By-Wire Feedback Design

To design the steer-by-wire functionality, first the control activity of the driver–vehicle interaction has to be explained. At slow vehicle velocities, the driver controls the vehicle path by the yaw velocity; feedback variable is the curve difference which means

■ **Table 11.3**
Significant steering functions

	Function	Quality
Improved vehicle agility and steering comfort	Variable steering torque	Velocity-adapted steering torque, <ul style="list-style-type: none">• high support for easy parking• low support for good response at high speed
	Variable steering ratio	Velocity-adapted steering ratio <ul style="list-style-type: none">• Less steering wheel angle at low speed maneuvers• Smooth handling at high speed
	Feed forward steering	<ul style="list-style-type: none">• Agile vehicle behavior at high steering wheel velocities by phase reduction between steering input and vehicle dynamics
Improved vehicle stability	Yaw rate control	Control of desired yaw rate in critical driving situations by: <ul style="list-style-type: none">• steering angle addition• steering torque variation
	Disturbance response control	<ul style="list-style-type: none">• Yaw torque compensation at μ-split conditions• Side wind compensation• Reduction of road disturbances or changed vehicle parameters (e.g., loading)
Driver assistant systems	Lane assistant	<ul style="list-style-type: none">• Course holding• Lane change assistant• Collision avoidance
	Autonomous steering	<ul style="list-style-type: none">• Parking assistant• Traffic jam assistant• Autonomous driving

the path difference in the foresight point. At higher velocities, the driver controls the lateral acceleration; feedback variable is the yaw velocity respectively the lateral velocity in the foresight point (see ● [Table 11.4](#)) (Huang 2004).

Many steer-by-wire functional design proposals can be found in the literature (Bünthe et al. 2002; Koch 2009; Odenthal et al. 2003). Next, two SbW functional design concepts are explained in detail. As explained in the introduction (chapter input module characteristics) the force feedback can be designed totally free with active input modules. In principal force input and displacement feedback, respectively, displacement input and force feedback are useful control strategies for a SbW application (Boller and Krüger 1978).

- Concept 1: Force input and displacement feedback

In a low-speed, steady state cornering situation, the steering torque (force input) is linked to the yaw velocity, the steering wheel angle (displacement feedback) is constant (steady state cornering means constant curve radius). Increasing velocity first means increasing steering torque and constant steering angle (because of steady state cornering). At higher velocities, the steering torque is equivalent to the lateral acceleration, the steering wheel angle increases linear to the velocity (because yaw velocity is proportional to vehicle velocity). This behavior is similar to variable steer ratio systems with direct ratio at low speed and indirect ratio at high velocities.

In a critical understeer situation, the driver's steering wheel angle demand is higher than the real steering wheel angle, defined by the low yaw rate. Therefore, the steering wheel becomes hard. In a critical oversteer situation, the vehicle yaw rate is higher than the driver's wheel angle demand and the steering wheel became weak. To stabilize the vehicle in this situation, the driver has to keep the actual steering wheel angle constant which is easy to handle.

At side wind disturbance when the vehicle begins to yaw, the driver will realize a certain steering torque. If the vehicle should go on straight ahead, the driver has to keep the required steering angle, the resulting steering torque means a lateral acceleration input against the disturbance what keeps the vehicle on course.

- Concept 2: Displacement input and force feedback

In a low-speed, steady state cornering situation, the steering wheel angle (displacement input) will increase proportional to the vehicle velocity at constant steering torque (force feedback). At higher velocities, the steering wheel angle will increase

■ **Table 11.4**

Driver's control variables for the vehicle lateral dynamic control

Velocity	Drivers demand	Feedback variable
Small (<30 mph)	Yaw velocity $\dot{\psi}_{\text{wanted}}$	Curve κ_{result}
Middle to fast (>30 mph)	Lateral acceleration $a_{y_{\text{wanted}}}$	Yaw velocity $\dot{\psi}_{\text{result}}$

disproportionately high with linear increasing steering torque; in this situation, Concept 2 is not as useful as Concept 1.

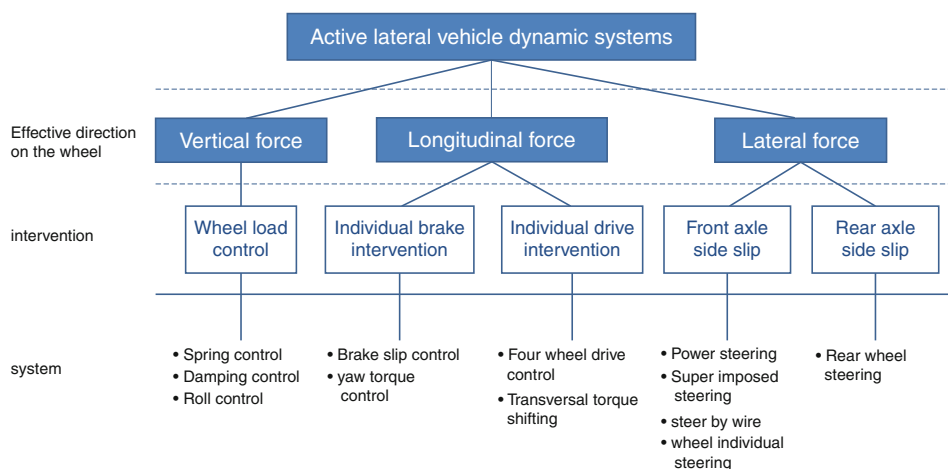
In a critical understeer situation, the steering torque decreases due to the fact that the resulting yaw velocity is lower than the wanted one. Intuitively the driver will follow this lower steering torque which means higher steering wheel angles; this is not useful in this situation because the front lateral tire force is on the limit. In critical oversteer situations, the steering torque will increase, the steering wheel becomes “harder,” different to the usual vehicle behavior in this situation.

At side wind disturbance, the vehicle begins to yaw, and equivalent to Concept 1 the driver realizes a certain steering torque. But different to Concept 1, the driver has to “follow” this torque to compensate the disturbance.

Summing up in a theoretical view, Concept 1 seems to be more useful than Concept 2. However, this consideration has to be calibrated in many driving and even in stop situations. To keep the costs in a limit, the feedback motor is as small as necessary. That means the motor is less strong than the driver, and to prevent endless steering by the driver, a mechanical stop has to be integrated in the concept. The control strategy and fade in of this mechanical stop defines the feedback quality of steer-by-wire, too.

3.3 Lateral Dynamic By-Wire Control Systems

► *Figure 11.16* shows active lateral vehicle dynamic systems, their effective direction on the wheel and their intervention variable. Beside this, some unorthodox systems like active aerodynamic, active camber, or active mass positioning are thinkable but not used.



■ Fig. 11.16

Active lateral vehicle dynamic systems

3.3.1 Vertical Force

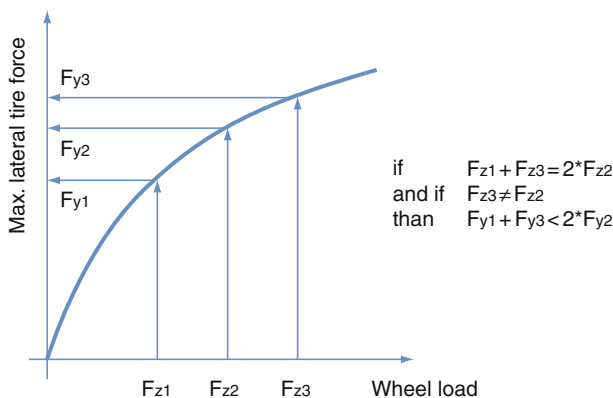
Wheel load control systems like spring, damping, or roll control are able to shift dynamic wheel load differences between front and rear axle. Since the lateral tire force depends nonlinear (digressive) from wheel load, the proportion of front and rear axle side force can be controlled by active vertical load systems.

► [Figure 11.17](#) shows a digressive tire characteristic. In general, optimal side force is given at the same wheel loads at all wheels. Since the center of gravity is above the street, the sum of the outer wheel loads is higher than the sum of the inner wheel loads. This correlation is just defined by center of gravity and vehicle mass. By implementing stabilizers, the ratio of outer front wheel to outer rear wheel, respectively, and inner rear wheel to inner front wheel can be adjusted. So, the wheel difference of outer and inner wheel at each axle can be adjusted; higher wheel load difference means lower side force potential. Therefore, most vehicles have stronger front stabilizers to prevent oversteer behavior in critical driving situations.

With vertical active systems, this functionality can be used to control the vehicle lateral dynamic; however, the influence of vertical systems to the lateral vehicle dynamic is smooth but low compared to longitudinal and lateral systems.

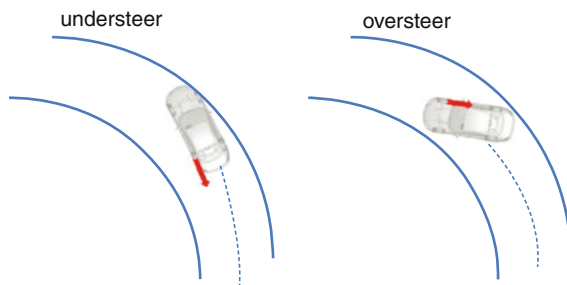
3.3.2 Longitudinal Force

Longitudinal force control systems (explained in chapter longitudinal dynamic) can be used to control the lateral dynamic too. Brake slip control (ABS) will prevent locked wheels which means the lateral tire force potential and therefore lateral vehicle stability is still given. If the difference of desired (calculated from steering wheel input and vehicle velocity) and measured yaw rate pass over a certain value, the yaw rate may be controlled by wheel-individual brake intervention (► [Fig. 11.18](#)).



■ Fig. 11.17

Nonlinear (digressive) tire characteristics



■ Fig. 11.18

Individual wheel brake intervention



■ Fig. 11.19

Four-wheel drive and torque shift system on the rear axle (BMW X6)

If the vehicle is in an understeer situation, the front wheels are at their lateral tire force limits and additional longitudinal force on the front wheels may amplify the understeer behavior. Therefore, the inner rear wheel has to produce the right yaw torque. If the vehicle has oversteer, the rear wheels are on their lateral tire force limit and therefore the outer front wheel has to control the yaw torque to keep the course.

The same principle can be used to control the vehicle dynamic by traction slip with longitudinal and/or transversal torque shifting systems (● Fig. 11.19).

The power divider (blue box in ● Fig. 11.19) controls the drive torque between front and rear axle and since the maximum wheel force is limited more longitudinal force means less lateral force on this axle. Additionally, at given drive torque on the rear axle, the yaw

torque can be controlled by the torque shift system (red box in ► Fig. 11.19). By the drive torque shift system, the torque ratio between left and right side can be controlled and therefore a certain active yaw torque can be brought in the vehicle dynamics.

However, brake intervention has a positive effect in critical situations by reducing the vehicle speed in general but should be used carefully at certain limits because the driver feels a discrete, noncomfortable intervention. Different from this drive torque shifting can be used smooth and therefore permanently both, in critical situations and to improve the vehicle agility.

3.3.3 Lateral Force

Lateral force control systems use the wheel side slip to control the vehicle. Unlike single lane vehicles such as motorbikes, the camber angle influence is secondary. To control the side slip, steering systems, especially on the front axle, are used. Next, different active side slip control systems from simple to complex are shown.

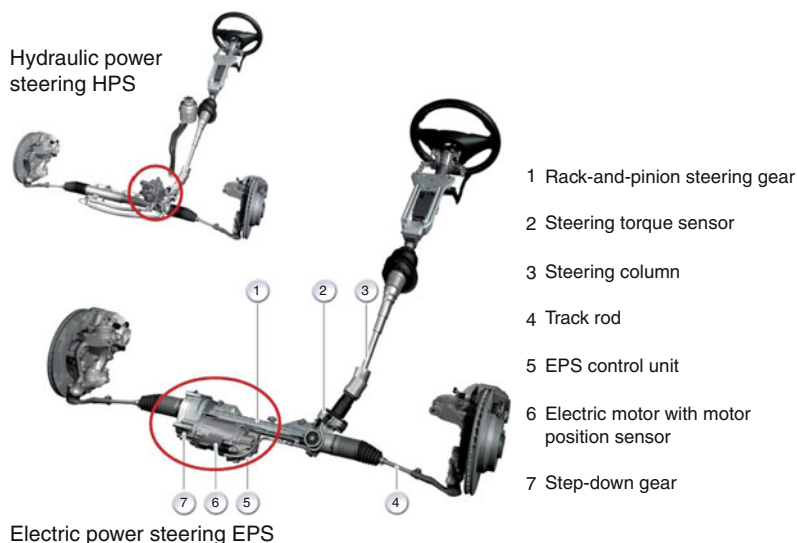
Power Steering Systems

Power steering systems support the driver reducing his required steering torque which is helpful to reduce the steering ratio and the steering wheel diameter. Hydraulic power steering (HPS) systems were first on market. Permanently hydraulic fluid flow is controlled by a torque rod in the steering column. This works very well and is cheap and reliable. However, fuel efficiency is not given due to the permanent fluid flow and innovative steer torque control functions are complex to integrate in hydraulic systems. Therefore, in the last years Electric Power Steering (EPS) systems came on the market. ► Figure 11.20 shows both concepts and the components of the EPS.

The EPS motor may be designed at the steering column, at the pinion, as so-called double pinion or axially parallel to the rack (► Fig. 11.6). The first systems are cheaper, the last ones stronger. Heavy vehicles with high front axle load need higher voltage systems (42 V) to limit the current demand.

With EPS, the steering torque can easily be controlled such as depending on the vehicle velocity. High torque support will be given at low speeds and for easy parking, low support will improve the response at high speed. The steering return ability can be improved by EPS as well as steering damping. Driver-independent steering torque can be used for assistant systems (lane keeping systems) as well as supporting the driver doing the right maneuver in critical situations. However, the front wheel slip angle is exactly defined by the driver's steering wheel angle.

Driver assistant systems up to autonomous driving can be realized by EPS too. First systems on market are course holding, lane change, collision avoidance and autonomic parking systems. Traffic jam systems, particularly, general autonomous driving is still part of research projects (Seewald 2008).



■ Fig. 11.20
Hydraulic and electric power steering system

Variable Steer Ratio

Variable steer ratio systems enable an individual steering ratio depending on steering wheel angle or vehicle velocity. Thus, the ratio can be optimized to the drive situation. In 2000, Honda launched the S2000 Type V equipped with the world's first electric power variable gear ratio steering (VGS) system (► Fig. 11.21). The steering ratio depends on the movement of lever B.

However, compared to superposition steering systems, the concept has a similar mechanical complexity by less functionality.

Superposition Steering

► Figure 11.22 shows two different superposition steering system concepts with by-wire functionality and mechanical fallback.

On the left side, a hydraulic system enables an additional move of the steering rack box, on the right side a worm gear system adds steering angles to the drivers input. In case of system failure in the hydraulic system, the worm gear system will be shut down; thus the fixed rack box, particularly, the fixed ring gear ensures steering functionality.

► Figure 11.23 shows the realization of superposition steering with worm gear.

► Figure 11.24 shows the variable steering ratio with respect to the vehicle velocity done by the superposition steering system. The constant basic mechanical steering ratio will be reduced at low-speed maneuvers and raised at high-speed maneuvers; thus comfort at low speed and stability at high speed are given without compromise. The feedback steering torque is influenced marginally by the superposition steering.

Different from variable steer ratio systems, the superposition system allows bringing in additional wheel angles to the driver ones.

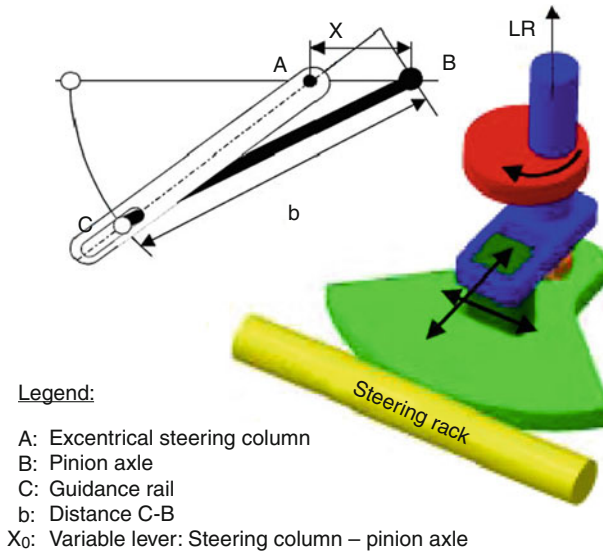


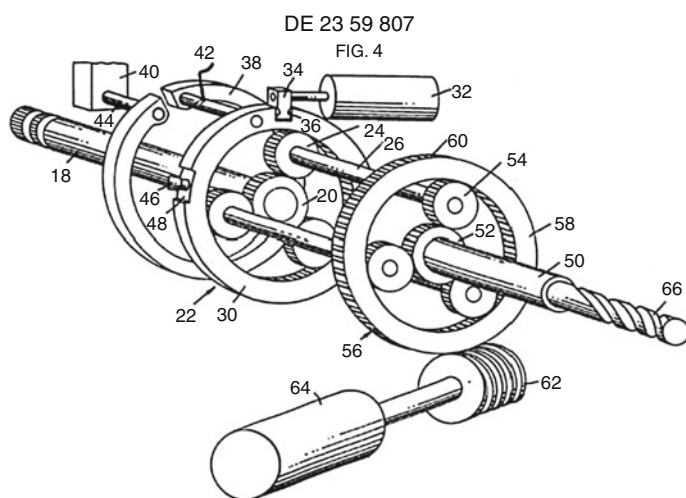
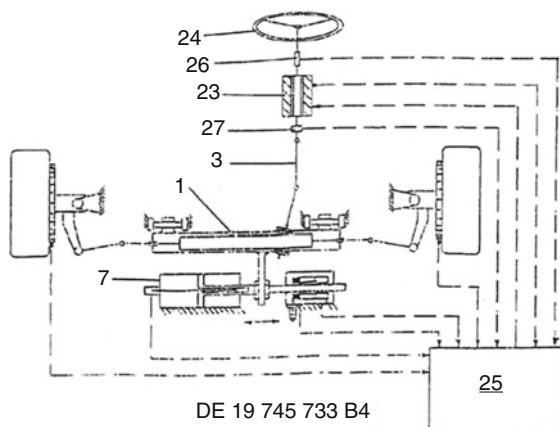
Fig. 11.21
VGS concept to vary the steer ratio

In **Fig. 11.25**, the principle function of agility improvement by feed-forward steering is shown. In conventional steering systems, the wheel angle follows the steer step input proportional with a given phase difference due to mechanical stiffness. The resulting yaw rate staggers again to the wheel angle due to given tire dynamics and overswings the steady state yaw rate because of the vehicle inertia. By adding a dynamic steering angle, depending on wheel angle and wheel angle velocity, the feed-forward wheel angle is no longer proportional to the steer input which results in a faster raising yaw rate and less overswinging in the steady state phase which means more vehicle agility.

With feedback control, superposition steering systems can be used to control the yaw rate by additional driver-independent wheel angles. Thus, yaw torque compensation at μ -split conditions can be done as well as side wind compensation and vehicle adaption to changed vehicle parameters. The vehicle dynamics can be controlled to prevent sliding. Since the intervention is smooth, wheel angle control systems will be used preliminary compared to wheel slip control systems. However, in understeer situations, the maximum lateral wheel force on the front axle is on the limit; therefore, vehicle dynamic stabilization with superposition steering works best in oversteer situations.

Rear Wheel Steering

Superposition steering systems use the same control variable as the driver, namely, the front wheel angle. In best case, these systems are as good as a perfect driver. Different to this, rear wheel steering systems can improve the vehicle behavior independent from the driver's input. The rear wheels' steering angle depends on different parameters such as front wheel steering angle, vehicle velocity, yaw velocity, or lateral acceleration.

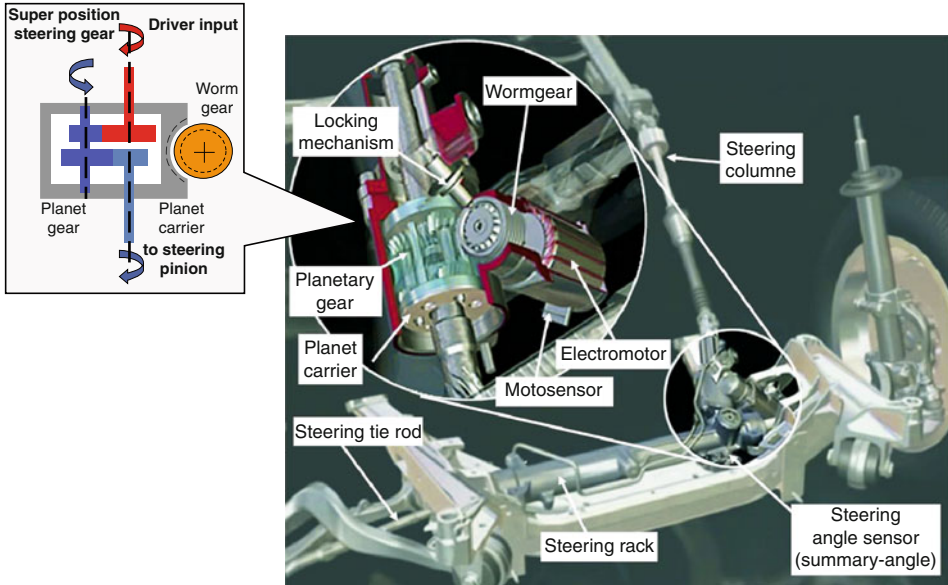


■ Fig. 11.22

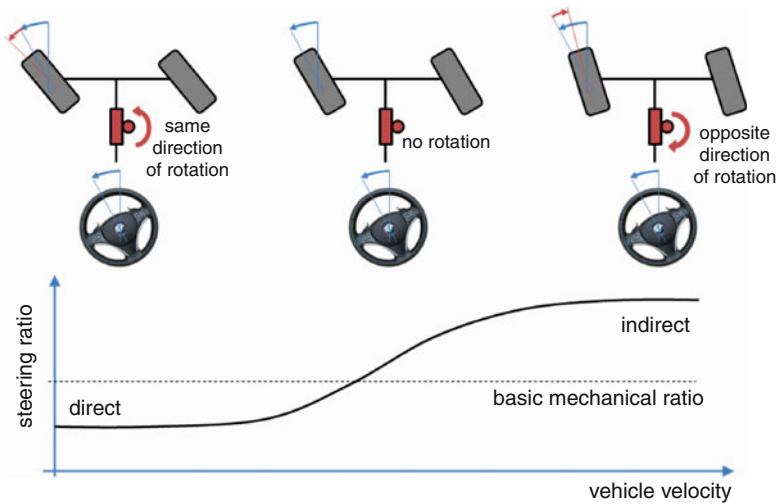
Two patents with concepts of possible superposition steering systems

► [Figure 11.26](#) shows the principal functionality of rear steering. At very low speed, the vehicle curve radius R can be found as shown in *a*. By using the rear wheel steering in opposite to the front wheel steering, the curve radius can be reduced which means the vehicle behavior is like a smaller vehicle without rear wheel steering, the vehicle became agile (► [Fig. 11.26b](#)). Using the rear wheel steering in the same direction as the front steering, the curve radius became bigger (► [Fig. 11.26c](#)) and the vehicle more stable (like a longer vehicle). In medium- and high-speed maneuvers, this phenomenon can be used to reduce the vehicle side slip angle (Pruckner 2001).

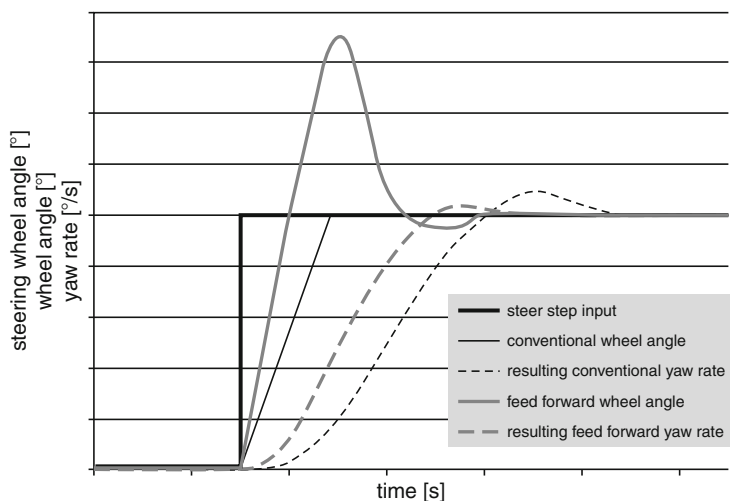
In 2010, the BMW 7 series placed the so-called integrated active steering which means the combination of superposition steering on the front axle and HSR (rear side-slip



■ Fig. 11.23
Active steering in the BMW 5-series (Fleck 2003)

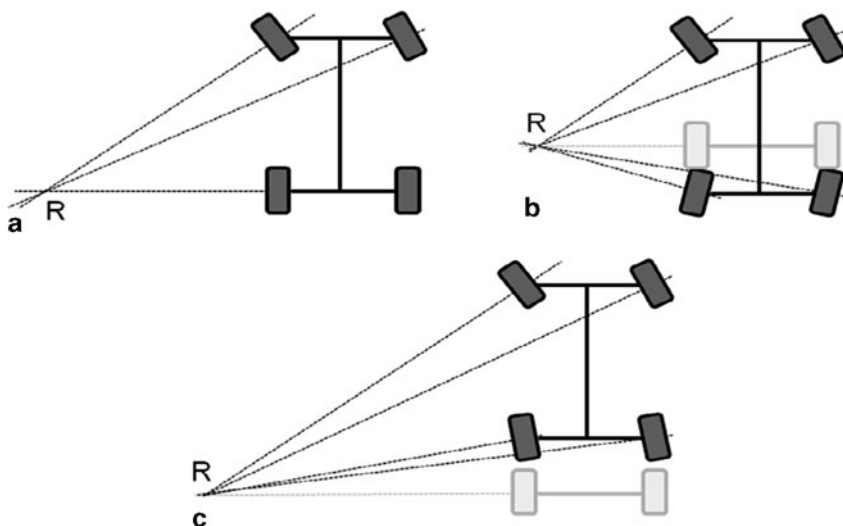


■ Fig. 11.24
Variable steering ratio by superposition steering system



■ Fig. 11.25

Feed forward steering to improve vehicle agility

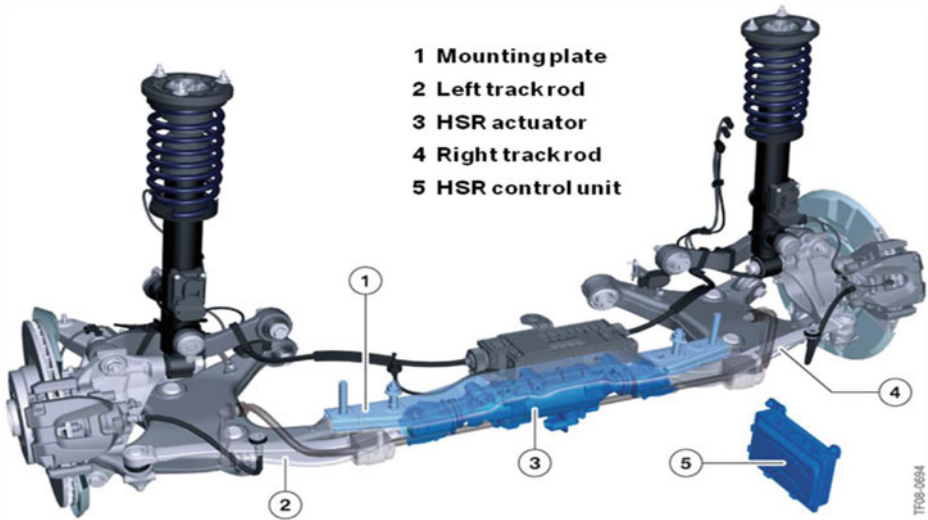


■ Fig. 11.26

Functionality of rear wheel steering systems

control) on the rear axle (Wallbrecher et al. 2008). Figure 11.27 shows the components of the HSR actuator in the rear axle.

Different to superposition steering, the vehicle dynamic can be controlled by rear wheel steering best in understeer situations. Therefore, the combination of both control



■ Fig. 11.27
Rear wheel steering system in the BMW 7 series (2010)

variables on front and rear wheels, respectively, is best to improve the vehicle agility and safety by lateral dynamic control systems.

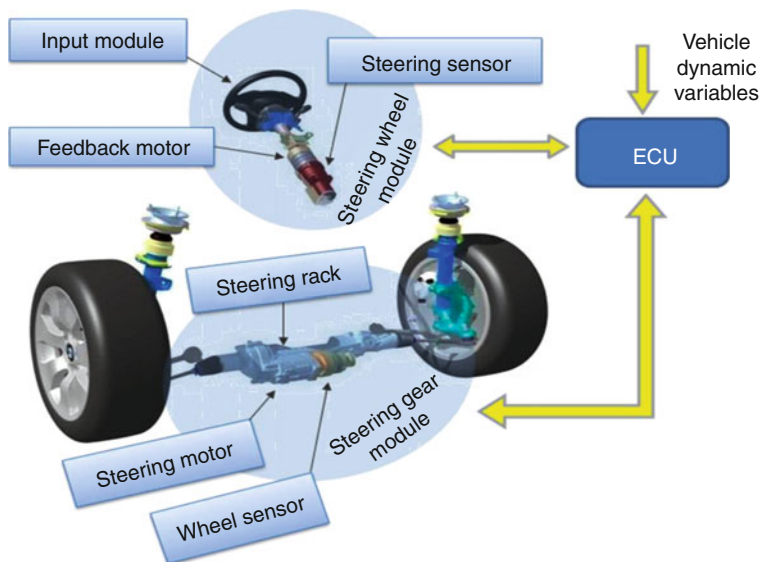
Steer-By-Wire

● [Figure 11.28](#) shows the components of a true steer-by-wire (SbW) system. The steering system is divided in steering wheel module and steering gear module connected just by an electronic control unit (ECU).

The steering wheel module is characterized by input module (e.g., steering wheel or stick), feedback motor, and sensors. The steering gear module contains steering motor, sensors, and steering rack. Both modules are connected by wire and controlled by an electronic control unit (ECU) which also uses general vehicle variables (e.g., velocities, accelerations, etc.). In comparison to superposition steering system combined with power steering, the following qualities of SbW can be defined.

Key benefits of SbW in comparison to superposition steering with power steering:

- In conventional vehicle designs, left and right steering options are foreseen even if there is only one steering column in a car. And therefore by the abstinence of the steering column, engine and exhaust design in the vehicle front can be done better without compromise.
- Additional to the package, the crash design can be improved since the steering column may not enter the passenger cap.
- Different to superposition steering systems, where a certain connection of wheel lateral force and steer torque is given, SbW enables a totally independent design of wheel angle and steering torque.



■ Fig. 11.28

Steer-by-wire components

Drawbacks of SbW in comparison to superposition steering with power steering:

- No additional useful steering functions can be done by true SbW.
- Since the driver input torque cannot be used, the power demand at input motor and steering motor is higher.
- The feedback torque realization is more complex.
- The required redundancy due to functional safety and availability means more electronic components and therefore higher cost and weight.

Especially to reduce cost and weight, a simple mechanical fallback system with hydraulic components (Heitzer and Seewald 2000) but without steering column can be done (► Fig. 11.29).

Usually the left valve is in the bypass position so that there is no pressure in the left hydraulic system. The steering rack is controlled just by the right-side hydraulic system. The valves in ► Fig. 11.29 are in fault position; that means the right hydraulic system is in bypass position and the steering rack is controlled by the hydraulic cylinder near the steering wheel.

However, since costs and complexity for SbW are higher than for superposition steering combined with power steering at given safety and reliability demand, SbW cannot be found in serial production vehicles today. The additional package and design area seems to be not big enough compared to cost and complexity.

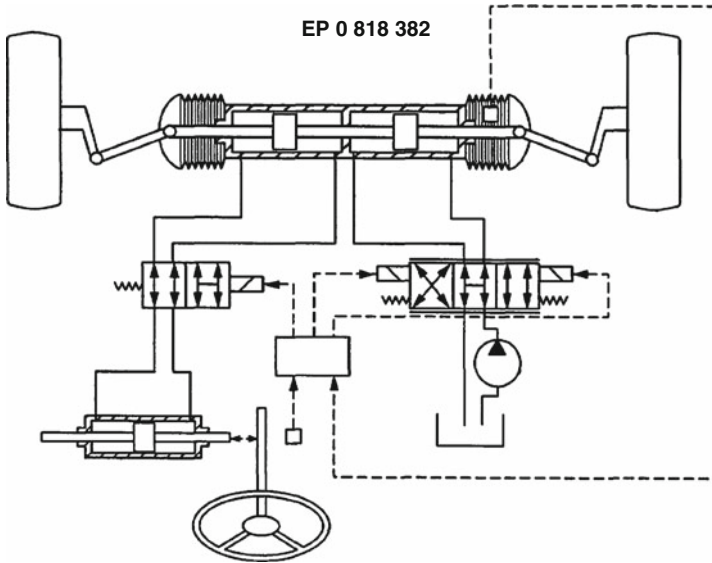


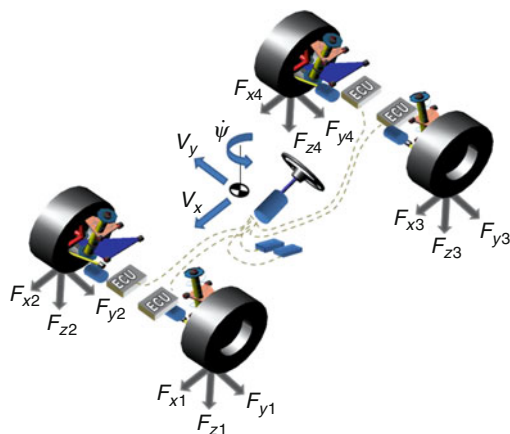
Fig. 11.29
Steer-by-wire with hydraulic fallback system (EP 0 818 382)

4 Integrated Vehicle Dynamic

Since longitudinal and lateral vehicle dynamic is influenced by each other, the above-described systems can be used for an integrated control approach. A highly integrated system explained next is the so-called corner module concept with four individual corners on a vehicle. Functional integration means integrated control strategy that is explained thereafter in this chapter.

4.1 Corner Module

Active vehicle dynamic systems are used more and more in standard vehicles. Besides aerodynamic effects, all four wheels bring forces into the vehicle. Each wheel has 6 degrees of freedom; that means 24 possible force input variables to control a vehicle. Assuming that wheel base and track is constant, four input variables (steering angle, camber angle, wheel spin and vertical wheel movement) at each wheel can be defined. Today, the camber angle will be controlled passively by the mechanical axle design in a way that the wheel is as flat as possible on the street in every driving situation. Active chamber wheel systems are very expensive and therefore not useful. That means a fully equipped by-wire vehicle enables three input variables per wheel (steering angle, wheel spin, and wheel load), respectively, 12 input variables into the vehicle. ➤ [Figure 11.30](#) shows the concept of four equivalent



■ Fig. 11.30

By-wire corner modules with 12 input variables and 3 control variables

corner modules controlled by wire. Each corner enables wheel spin (drive or brake), steering and wheel load control.

The global horizontal vehicle movement can be defined by three vehicle parameters which are longitudinal velocity V_x , lateral velocity V_y , and yaw rate $\dot{\psi}$ (see Fig. 11.30). By these three variables, the plain vehicle state (longitudinal and lateral) can be defined exact. Three control variables and 16 input variables means the system is highly underdetermined, for example, deceleration can be done by spin reduction as well as by opposite steering angles at one axle (like a snow plow). On the other side, the lateral velocity and the yaw rate may be controlled by all 16 input variables as explained above.

In a vehicle dynamic view, single wheel steering makes no sense because in high-speed cornering maneuvers, the mean dynamic influence is given by the outer wheel due to the wheel load. Additional, the control strategy for just a simple thing like going on straight ahead what means an exact synchronization of left and right wheel is very complex in comparison to a simple mechanical steering rack. Of course, due to the underdetermined system configuration, the safety concept in case of system failure can use the redundancy, but the general vehicle dynamics cannot be improved by more input variables. Four-wheel drive is just useful in snowy mountains conditions and high rear wheel steering angles means a big wheel housing and therefore less customer space. All in all, a high integrated corner module concept is not useful in view of costs and vehicle dynamics.

4.2 Control Strategy

Many active vehicle dynamic systems mean many input variables to control the vehicle. The following vehicle dynamic control systems are state of the art:

- Antilock brake system
- Dynamic stability control

- Active four-wheel drive
- Drive torque shifting
- Superposition steering
- Rear wheel steering
- Vertical damping control
- Active roll control

Every system may influence the horizontal vehicle dynamic, which can be defined by the three state variables longitudinal velocity V_x , lateral velocity V_y , and yaw rate $\dot{\psi}$ (see Fig. 11.30). Beside the fact that each control system has to do the right job for itself, a good tuning of all systems together has to be done as well which becomes very complex with an increasing number of systems.

An Integrated Chassis Management System has to take into consideration the individual system configuration required by the customer. Therefore, a daisy chain concept is defined where smooth input variables like wheel angles are used first for permanent and safety relevant systems like anti-wheel lock later but dominant in critical driving situations.

Figure 11.31 shows the daisy chain principle. The yaw rate control delivers a torque demand to actuator 1. If the acting variable U_1 will achieve the wanted yaw velocity, the control task is finished. Otherwise, if the difference in wanted and measured yaw-rate is still given, the next actuator 2 is used and so on. This cascaded principle depends on maneuver and vehicle dynamic behavior and is controlled by the state-dependent rule.

The adjustment has to make sure that more active systems mean more customer benefit and even same or better safety level. This means a high development effort to provide right functionality in every possible driving situation.

Taking into consideration that more active system (e.g., energy recuperation at different vehicle axle) will increase the development demand, new control strategies like control allocation systems has to be found (Knobel 2009). In the airplane industry, especially ultrasonic planes, these kinds of control systems are state of the art (many input variables have to control some plane state variables).

Main idea of control allocation is a vehicle model controller with all input, state, and output variables. The influence of every single input in combination with the other ones will be pre-estimated by the controller, the underdetermination allows to define additional criteria like highest safety, much driving pleasure, or high economy.

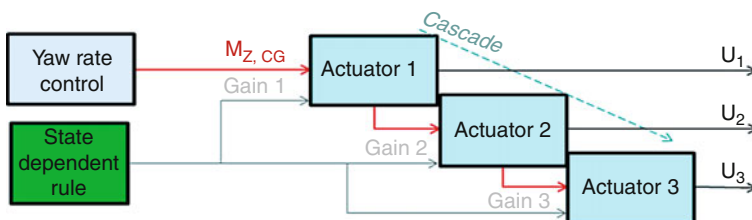
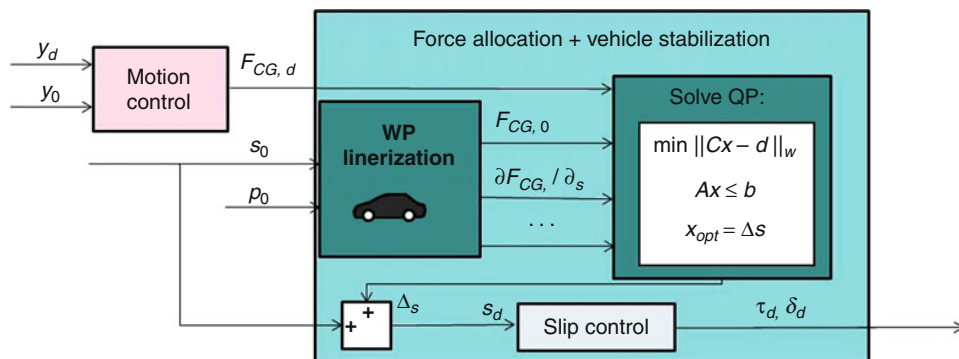


Fig. 11.31
Schematic of a control concept with cascaded daisy chain principle



■ Fig. 11.32

Schematic of a control allocation algorithm

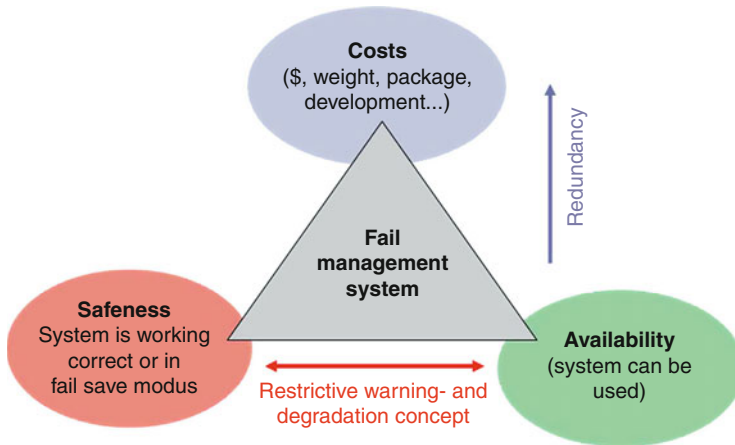
In [Fig. 11.32](#), a control allocation principle is shown. First, the motion control block delivers the necessary forces to bring the measured variables (y_0) to the delivered ones (y_d). The linearized vehicle model delivers the actual vehicle forces and states as well as their derivations about the acting variables (longitudinal and lateral tire slip). The Solve QP block estimates the optimal acting variables and their interaction depending on system availability, system dynamic, and the above explained additional criteria. At least, these optimal acting variables, longitudinal and lateral tire slip values, have to be regulated by the slip controller.

The complex control algorithm has to be found ones for a vehicle with all control systems integrated. If the vehicle is equipped with less systems or a failure is detected online, just some lines in the Solver has to be multiplied by zero (Krueger 2010).

5 Functional Safety and Availability

As mentioned already in the previous chapters, one major topic is the functional safety of safety critical systems such as by-wire systems. In general, by-wire systems do have the ability to give actuation commands which are at least as fast and strong as a good driver, which means that any unwanted command could lead to major longitudinal or lateral discrepancies resulting fatal accidents. Moreover, the availability of systems without a mechanical fallback linkage which do not represent a safe state when failed needs to be taken into consideration when designing such a system. Hence, steer-by-wire and brake-by-wire systems without mechanical backup must be realized in a fail functional matter, in contrast to some of the drive-by-wire functionality as well as driver assistance systems which are normally fail silent.

Additionally, the customer expects all functions to be fully available at all times. [Fig. 11.33](#) shows the trade-off between safeness, availability, and costs whereby costs are not only the costs of redundant parts, but also the development costs and time as well as the package integration and system weight.



■ Fig. 11.33

Trade-off between safeness, availability and costs in a fail management system

A system that must be fail functional will still have convenient functions which will be realized as fail silent. A brake-by-wire system, for example, will need to keep the ability to reduce the speed of the car after the first system failure, but could well have degradation in higher level functions such as stability brake actions or cruise control with braking, which will be in a safe state when switches off. The steer-by-wire feedback motor is important for a good feeling, however, if the motor fails, a simple fall back system like a passive spring satisfies low dynamic requirements. In that sense, a good warning and degradation concept helps to increase safeness and availability by reduced functionality and therefore save costs.

5.1 Basic Design of Safety Critical Systems

Functional safety means that the system must behave in all situations. It has to be possible for the driver to control the vehicle and bring it to a safe position. In order to design a safety critical system, one has to know all system influences to the vehicle and the safe state. First, the actuation and its restrictions must be known, from which one can derive the reaction of the vehicle. This is essential to estimate if a driver is able to handle the failure reaction of the vehicle and keep it in a safe state. Very often, this depends on the force and the speed of the actuator. After that, the vehicle reaction must be analyzed and the maximum time a given failure is present has to be derived to learn the maximum failure latency time of this event. The last step is to design a function and/or system degradation concept and especially for safety critical by-wire systems a detailed fallback scenario. As mentioned above, some functions or systems can just be switched off and the vehicle drives just fine, and others are essential to control the vehicle and must not be switched off completely. On that account, the safe state of the first ones can be designed as fail silent, hence once a failure has been recognized they are switched off and therefore silent. The latter systems must be designed as fail functional, meaning, once a failure has been

detected, the system has to adjust its functionality in a way that the basic function is still given. Of course, the degradation of the system also needs to be shown to the driver to have a vehicle serviced and/or to take precautions that another failure does not cause fatal vehicle reactions.

The three essential steps of designing a safety critical system are described below, in a little more detail. However, when designing a safety critical system, such as a drive-by-wire system, one should take the relevant regulations into considerations, especially helpful are: European homologation regulations for brake systems ([ECE R 13](#)), the international standard for industry: Functional safety of electrical/electronic/programmable electronic safety-related systems ([IEC 61 508/EN61508](#)), and the new standard under development, adjusted for automotive applications: Road vehicles – Functional safety ([ISO 26262](#)).

5.1.1 Safe State of a Vehicle with the System(s)

First of all, the system and the vehicle reaction must be known in the way that the safe state of this system can be distinguished, meaning what is the best strategy and actuator position in case of a detected failure of the system to do no harm. Very often, this is not easy to tell, and one has to consider the state of the art. For example, one could consider the safe state of a passenger vehicle being stand still. For most cases this is true, but if the position of the vehicle is on a rail crossing, this might not be the safest state. So, the safe state might be a vehicle that can be moved, even very slowly. However, that would mean that the combustion engine must be redundant, in order to be able to move the vehicle in case of an engine failure. This is where the state of the art comes in. Normally, no vehicle is equipped with more than one engine. However, thinking of hybrid vehicle, this could be changed and in case of a failure in the combustion engine, a hybrid vehicle might be able to move for a short time.

As described in the previous sections, one distinguishes between fail silent and fail functional. A fail silent system is a system which has its safe state when it is not functioning, therefore silent. For example, a cruise control system is in a safe state when it does nothing. The driver can control the speed and distance to the previous vehicle on his or her own. A brake system should have some kind of function when failing. Therefore, hydraulic brake systems are equipped with two separate hydraulic circuits. So the safe state is a brake system which does not brake unintentionally, but meets the regulations in deceleration and brake pedal force when the driver wants to decelerate.

5.1.2 Danger and Risk Analysis

Next, not only the complete system, but also all functions within the system have to be analyzed, which degree of harm they could do. For that, one has to know the impact of the system and all the functions on the vehicle in all cases, meaning also in the case of any possible malfunction. Any undesired actuation leads to a response of the system and therefore to an impact on the driver or the vehicle. The aspects of malfunctioning include

not only functioning or not, but also actuating the actuator too much, too little, too early, too late, with a too high gradient, etc.

Also, the question whether this response can be controlled easily by the driver has to be considered. For this, it is very often necessary to consider also the situation in which the vehicle is. Especially with vehicle chassis systems, the vehicle reaction is largely different depending on the road friction coefficient, the speed of the vehicle, and the vehicle dynamics. For assistance systems, often the driving situations are important, like being in a construction zone with reduced traffic lane, or in heavy traffic with little distance to the other vehicles.

Then, the controllability of the situation has to be evaluated. If the system has a malfunction, can the driver still control the vehicle and bring it to a safe driving condition? Does the system have to be disabled for this event, or can the driver overrule the system? This leads us again to the question if it is a fail silent system. If the system cannot be overruled, the question is, how long the malfunction can be present before the system must be switched off to be silent, or even the incorrect actuation being reversed. This time is known to be the failure latency time and leads to a major requirement of the degradation concept of the system.

The last step in the harness table is the evaluation of the harm, in case of the malfunction with all its consequences. This ranges from minor damages to fatally accidents. When malfunction, response, and driving situations are described, the possible consequences have to be evaluated.

All the above evaluations are very subjective. For that reason, the above-mentioned standards and regulations give hints on how to evaluate these cases and also include very good guidelines for designing, evaluation, and testing of a safety critical system within an automotive application.

5.1.3 Degradation

Having a system which cannot be designed as completely fail silent, one has to consider which functionality it has to provide, once the main function is not operational correctly. Of course, fail silent is a form of degradation that normally results in a display of the malfunction to the driver. More important are degradations of systems which the driver cannot overrule, and are essential for the driving task, for example, steering or braking with a system that can overrule the driver's input. Especially, steer-by-wire and brake-by-wire do have the ability to control the vehicle and to overrule the driver. These systems need the main functionality, that is, steering or braking according to the driver's input, even in the case of a system failure. For this reason, degradation leads to redundancy. A by-wire function might be able to be fail silent, if a mechanical backup can control the vehicle. Otherwise, it must be designed as fail functional, where a backup system can still control the main functionality. This is very often realized with the help of limitation by other systems. If for example a brake system does no longer have its full functionality, it is wise to limit the maximum speed of the vehicle. Also, the driver must be informed of the degradation in order to take precautions.

5.1.4 Environment for Safe Drive-By-Wire

One other aspect of safety critical systems is the availability of information necessary for its functions. Today's vehicles frequently use Controller Area Network (CAN) and Local Interconnect Network (LIN) for communicating instructions and receiving diagnostic information. True drive-by-wire architectures will require a safety critical data bus with inherent fault tolerance and higher bandwidth. Currently, two architectures, Time Triggered Protocol (TTP) and FlexRay, are being considered by automakers. The time-triggered technology in both of these protocols is viewed as essential to ensure that important messages always get through on the data bus at the right time.

5.2 Example: Electronic Throttle

Here, a short, not extensive, and full example of the safety concept of a typical electronic throttle is given. It shows the result of the above taken analysis, which is not shown here.

For safety reasons, the accelerator pedal sensor is equipped with two resistors (potentiometers) with varying operating ranges (for example 1–4 and 0.5–2 V) and separate circuits. The throttle valve is operated by a servomotor, and in addition, is monitored by two, mostly counter-rotating potentiometers. The engine control device evaluates these and many other signals and controls the servomotor on the throttle valve accordingly. It ought to be quite clear that the processing in a system, which is relevant for the safety, has an identical backup operation.

With the introduction of the electronic throttle, the emergency running function sometimes provided the grounds for discussions. If only one accelerator pedal sensor fails, an error would probably be noted and the car would carry on functioning. Should both fail to give a signal, it cannot react otherwise, than to distinctly raise the RPMs, either in certain situations or constantly, thus making it possible to reach the workshop. Modern electric throttle regulating takes into consideration the inlet manifold pressure or the air-mass, the clutch and brake light switch, and the wheel RPMs or the driving speed. Then, at the traffic lights, with the brake pedal pressed, the RPMs can be reduced and when pulling off or on an uphill stretch accordingly increased. However, it is important to know that a defective accelerator pedal transmitter cannot be effectively replaced, even by the most skillfully programmed emergency running feature. How then, should the control device know what the driver wants? The actuating of more than 50% of the load when emergency running, is neither possible nor does it make sense.

Additional the servomotor for the RPMs can also fail. In this case, the emergency running function is simpler, through a spring it takes on the exact position necessary for increased idling RPMs.

Much more precautions, like in the event of undervoltage or communication errors are taken, which are not explained here. Finally, it can be stated that the safety concept of a by-wire system should be taken into consideration right from the beginning of the design of the system which ought to be integrated in the vehicle.

6 Conclusion

Nowadays by-wire systems are used more and more in vehicle systems. Electronic throttle is state of the art as well as by-wire steering and braking functionality. However, true by-wire systems without mechanical fallback are not yet used in safe critical tasks like steering and braking.

By-wire system needs actuation with a speed of at least a very good driver, and force or capability of conventional vehicle systems. Due to safety requirements, the systems have to be designed redundant that means at least two or more sensors and actuators for one task. On the other hand, today's steering and braking systems enables steer-by-wire functionality with mechanical linkage which means less cost and complexity. Therefore, the benefits in vehicle design and additional customer space will, at the moment, not legitimate the additional by-wire costs.

References

Literature

- Boller HE, Krüger W (1978) Untersuchung eines Bedienelements mit Krafteingabe und Wegrückmeldung bei der manuellen Lenkung von Unterwasserfahrzeugen. *Z Arbeitswissenschaften* 32:254–260
- Breuer B, Bill K-H (2003) *Bremsenhandbuch*. Vieweg, Wiesbaden
- Bünte T, Odenthal D, Aksun-Güvenç B, Güvenç L (2002) Robust vehicle steering control design based on the disturbance observer. *Annu Rev Control* 26:139–149
- Continental Media Center. Automotive chassis products, electric hydraulic combi brake. <http://mediacenter.conti-online.com>. Accessed 19 Oct 2010
- Conti Homepage – SBA. Automotive chassis products, regenerative brake system. http://www.conti-online.com/generator/www/de/en/continental/automotive/themes/passenger_cars/chassis_safety/ebs/extended_functions/brems_systeme_en.html. Accessed 12 Nov 2010
- Eberl T, Stroph R, Pruckner A (2011) Analyse unterschiedlicher Bedienkonzepte der Fahrzeuglängsführung bei Elektromobilität, 2. Automobiltechnisches Kolloquium 2011, München, 11–12 Apr 2011
- ECE-R 13, Brake System Homologation. United Nations Economic Commission for Europe. <http://www.unece.org/trans/welcome.html>. Accessed 3 Nov 2010
- Eckstein L (2001) Entwicklung und Überprüfung eines Bedienkonzepts und von Algorithmen zum Fahren eines Kraftfahrzeugs mit aktiven Sidesticks, vol 471, *Fortschr.-Ber. VDI, Reihe 12*. VDI, Düsseldorf
- Fischle, Stoll, Heinrichs (2002) Bremsen auf höchsten Niveau- Die Sensotronic Brake Control. In: Die neue Mercedes-Benz E-Klasse, ATZ/MTZ Sonderausgabe. Vieweg, Wiesbaden
- Fleck R, Henneckke D, Pauly A (2001) Das steer-by-wire-system der BMW-Group zur Optimierung von Lenkkomfort, Fahrzeugagilität und -stabilität, Haus der Technik, Essen, 3/4 Apr 2001
- Fleck R (2003) Methodische Entwicklung mechatronischer Lenksysteme mit Steer-by-Wire Funktionalität. Tagung “fahrwerk.tech,” Garching
- Heitzer H-D, Seewald A (2000) Technische Lösungen für Steer-by-Wire Lenksysteme. Aachener Kolloquium, Aachen, Okt 2000
- Heitzer H-D, Seewald A (2004) Development of a fault tolerant, steer-by-wire steering system. SAE Nr. 2004-21-0046
- Huang P (2004) Regelkonzepte zur Fahrzeugführung unter Einbeziehung der Bedienelementeigenschaften. Dissertation, Fakultät für Maschinenwesen, TU München
- IEC 61 508. International standard for functional safety. <http://www.iec.ch/functionalsafety/>. Accessed 9 Nov 2010
- ISO 26262. International Organization for Standardization, Standards under development ISO 26262

- Part 1 – Part 10. <http://www.iso.org>. Accessed 12 Nov 2010
- Kilgenstein P (2002) Heutige und zukünftige Lenksysteme. Tag des Fahrwerks, Institut für Kraftfahrwesen, Aachen
- Knobel Ch (2009) Optimal Control Allocation for Road Vehicle Dynamics using Wheel Steer Angles, Brake/Drive Torques, Wheel Loads and Camber Angles, vol 696, VDI Reihe 12. VDI, Düsseldorf
- Koch T (2009) Bewertung des Lenkgefühls in einem Sportfahrzeug mit Steer-by-Wire Lenksystem. In: Aachener Kolloquium 2009, Aachen
- Krueger J (2010) Control Allocation für Straßenfahrzeuge -ein systemunabhängiger Ansatz eines integrierten Fahrdynamikreglers. In: Aachener Kolloquium 2010, Aachen
- Mueller J (2010) Active toe-compensation for by-wire steering systems. In: Tenth international symposium on advanced vehicle control, Loughbotough, Aug 2010
- Odenthal D, Bunte T, Heitzer H-D, Eivker Ch (2003) Übertragung des Lenkgefühls einer Servo-Lenkung auf Steer-by-Wire. Automatisierungstechnik 51(7):329
- Pruckner A (2001) Nichtlineare Fahrzustandsbeobachtung und – regelung einer PKW-Hinterradlenkung, D82. Dissertation RWTH Aachen, Institut für Kraftfahrwesen Aachen
- Prickarz H, Bildstein M (2009) Bremsanlage mit einer Vorrichtung zur optimalen Bremsdosierung. Bosch patent, Offenlegungsschrift DE102008012636
- Seewald A (2008) Auf dem Weg zur elektronischen Deichsel. Automobil-Elektronik Dez 2008
- Teitzer M, Stroph R, Pruckner A (2010) Simulation of an anti-lock braking system with electric motors during regenerative braking in powerful BEVs. Chassis Tech Plus, Munich, 8–9 June 2010
- Wallbrecher M, Schuster M, Herold P (2008) Das neue Lenksystem von BMW – Die Integral Aktivlenkung. Eine Synthese aus Agilität und Souveränität. In: Aachener Kolloquium 2008, Aachen
- Winner H, Isermann R, Hanselka H, Schürr A (2004) Wann kommt by-Wire auch für Bremse und Lenkung? VDI-Bericht 1828, Autorex 2004
- Zell A et al (2010) Active accelerator pedal as interface to driver. ATZ worldwide eMagazines edition Apr 2010, Accessed 18 Nov 2010

12 Energy and Powertrain Systems in Intelligent Automobiles

Ernst Pucher · Luis Cachón · Wolfgang Hable

Institute for Powertrains and Automotive Technology, Vienna
University of Technology, Vienna, Austria

1	<i>Introduction</i>	284
2	<i>Research Methodology</i>	285
2.1	Real-World In-Car Measurements	285
2.2	Calculation Model	288
2.3	Representative Real-World Driving Routes	289
2.4	Power Requirement of the Vehicle	291
3	<i>Energy Storage Systems</i>	293
4	<i>Vehicle and Powertrain Concepts</i>	294
4.1	Motivation for New Propulsion Systems	294
4.2	Internal Combustion Engine Vehicle	295
4.3	Hybrid Electric Vehicle	296
4.4	Fuel Cell Electric Vehicle	298
4.5	Battery Electric Vehicle	299
5	<i>Energy Consumption</i>	301
5.1	Tank-To-Wheel Energy Consumption	301
5.2	Primary Energy Consumption	303
6	<i>Summary</i>	304
7	<i>Symbols and Abbreviations</i>	305

Abstract: This chapter focuses on a methodological comparison of the energy consumption of passenger cars with conventional internal combustion engine and new electric powertrains.

The scientific methodology of the presented knowledge relies on simulation models validated by real-world tank-to-wheel emission and energy measurements performed in urban and extra urban areas as well as on freeways.

Furthermore, an overview of the recent development of energy storage systems, energy converters, and powertrains in developed regions is given.

Based on a statistically average European passenger car, fuel and energy consumption of conventional diesel, gasoline hybrid, fuel cell, and battery electric powertrains are compared. Special focus is given to the real-world energy consumption of the vehicles. This means auxiliary power units of the car such as onboard electronics, safety systems, heating, and air conditioning have significant influence on the fuel consumption, especially in driving conditions with lower speeds.

The results show that for typical real-world operational conditions during the seasons of a whole year the hybrid vehicle and, with some limitations in the range of operation, fuel cell vehicle are comparable to the usability of a conventional vehicle. Battery electric vehicles (BEV) have very limited range.

For the consistent comparison of propulsion systems with such different energy carriers, like liquid fuels, gaseous hydrogen, and electric power, the energy consumption with respect to primary energy supply was carried out. Based on this approach, the distance related energy consumption of hybrid and fuel cell vehicles is somewhat favorable compared to the conventional vehicle. Battery electric vehicles consume significantly more energy due to inefficiencies in power production and transmission as well as charging and discharging losses.

Nevertheless, cars with fuel cell and battery electric propulsion systems have the considerable benefit of local emissions-free driving.

1 Introduction

In recent years, public awareness of climate change and the CO₂ issue as well as matters of local air quality has increased dramatically. The automotive industry is especially affected by this discussion. Sustainable reductions of emissions in addition to the assurance of a secure energy supply are the drivers for the development of new powertrain technologies. The Californian zero emission regulations and the European CO₂ limits are additional driving forces toward zero emission propulsion systems.


Rating the efficiency of a vehicle can be a complex undertaking. Usually the type approval consumption is used as an indicator. Here a vehicle's tank-to-wheel consumption is determined by a defined driving cycle (e.g., NEDC), which is carried out on a test bed at stringently monitored conditions. However, real-world driving and operating conditions can differ seriously from the conditions used for the type approval.

For instance the important influence of auxiliary systems like air conditioning and heating are excluded. Their effect can be determined in real-world consumption and exhaust emission measurements of similar vehicles. Valuable data can be collected through the measurement of a vehicle's performance in real traffic conditions with traffic lights, traffic jams, uphill and downhill passages, and varying weather conditions. They can be used to improve powertrain simulation models and incorporate the impact of auxiliary power units within the vehicle at real-world driving and ambient conditions. The simulation of different driving pattern and therefore the analysis of different loads on the powertrain give a detailed overview of vehicle efficiency and energy consumption.

Furthermore, it is necessary to look into where the energy in the tank is coming from.

2 Research Methodology

A new comparison method for energy consumption of passenger cars with conventional internal combustion engines and cars equipped with electric powertrains will be presented. Tank-to-wheel simulations and evaluations were carried out mapping real driving conditions in extra urban, suburban, and urban areas. Furthermore, the consumption of the investigated vehicle concepts was calculated based on a typical primary energy composition of a developed region with a share of approximately 25% nonfossil energy sources.

In preceding research projects, complete powertrain data and real-world measurements of the compared vehicles have been analyzed. The measurements were conducted using a portable emission measurement system (PEMS) developed by the research group (Pucher et al. 2008a). Firstly, the measurement equipment was applied to evaluate the reference driving routes. Due to high temporal resolution of a complete dataset per second in combination with GPS tracking, the data can be used to reconstruct the reference driving routes for a simulation. The powertrains were analyzed globally, which include the interaction of all auxiliary systems and aggregates at real traffic conditions. The gathered data were evaluated and analyzed to plot the energy consumption versus the vehicle velocity and traction force. From these calculations, analytical functions of the energy consumption could be derived. These analytical functions in addition to the specifications of the vehicles used for comparison, such as dimensions, curb weight, resistance coefficients, efficiencies, and ambient conditions were defined as parameters and implemented in the simulation model. The calculations could be executed for a standardized driving cycle as well as for any other real-world driving route, including off-road sections.  *Figure 12.1* shows the methodology of the investigations.

2.1 Real-World In-Car Measurements

Measurements in real traffic conditions reveal the actual performance of a vehicle and are therefore of improved quality versus test bed measurements at controlled conditions

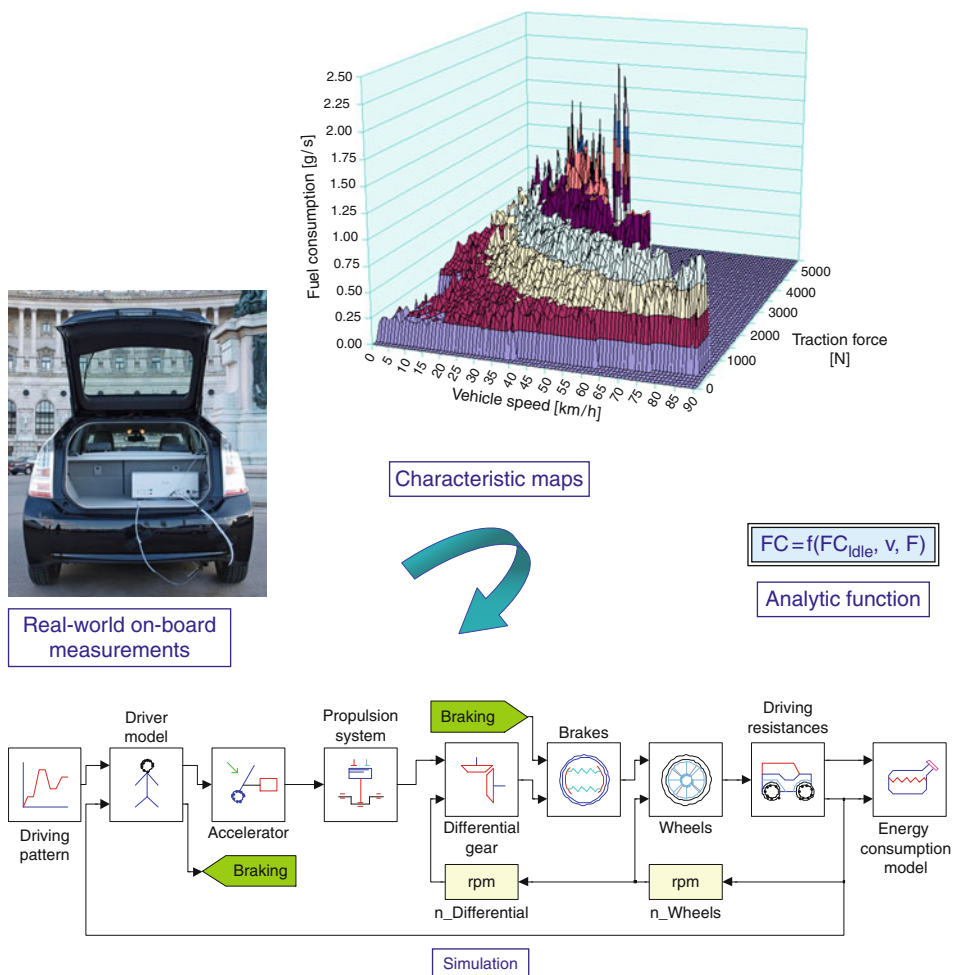


Fig. 12.1
Scheme of the methodology

(Pucher et al. 2008a). In particular, the effect of auxiliary consumers like air conditioning and heating system can be validated in real-world measurements.

The ultra-compact PEMS (Fig. 12.2) with a total weight of less than 15 kg measures the emitted exhaust gas mass emissions of the vehicle in real time without affecting the powertrain and vehicle characteristics. The gas concentration measurement works on the basis of a constant partial flow, which is taken from the exhaust gas in the tailpipe. This technology allows real-world in-use testing and provides more detailed data than conventional laboratories or vehicle test cells. The PEMS integrates accurate gas analyzers, exhaust mass flow meter, connections to the vehicle ECU, and Global Positioning System (GPS).

Emission components measured by the PEMS are CO₂, CO, HC, NO_x, and O₂. The CO₂, CO, and total HC concentrations were determined with nondispersive infrared



Fig. 12.2
Portable emission measurement system in a hybrid vehicle

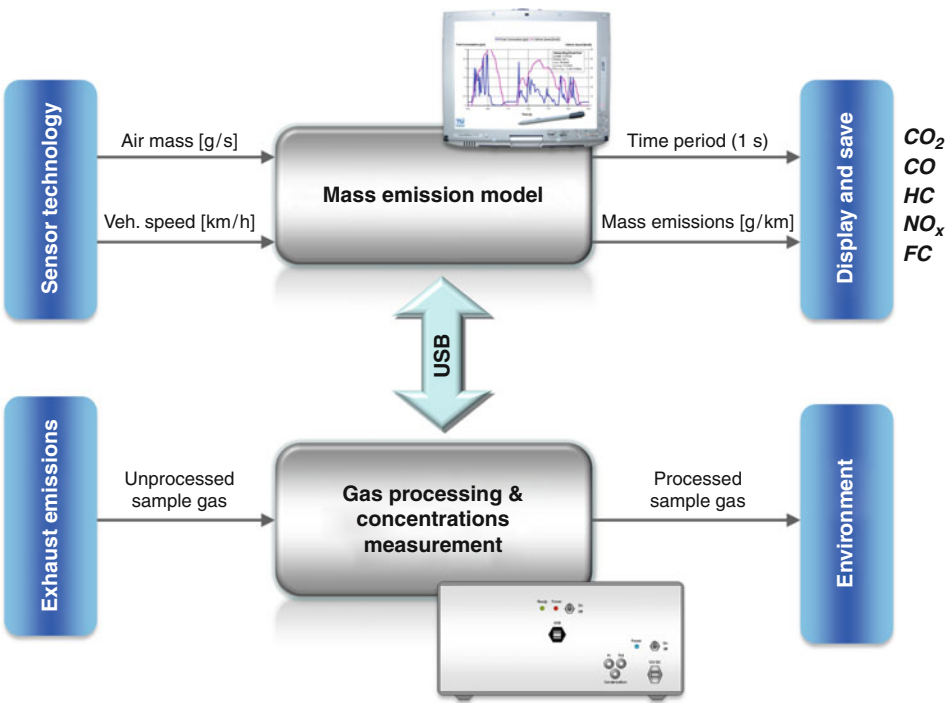


Fig. 12.3
Flow diagram of the real-world in-car measurement system

absorption (NDIR). Combined with the intake air mass flow and the vehicle speed, mass emissions of all exhaust gas components are calculated by the application of a chemical reaction model (► Fig. 12.3).

The PEMS provides the following output data:

- Air–fuel ratio
- Exhaust gas mass flow
- Exhaust gas mass emissions of the measured components in [g/s] and [g/km]
- Fuel consumption in [g/s] or [Liter/100 km] by means of carbon balance method

The results can be observed in real time on the control subnotebook in the vehicle. For subsequent analysis, every second a complete data string is saved.

Subsequently, a mechanical and electrical simulation model of the vehicles was created (Simic and Pucher 2009; Paces 2007).

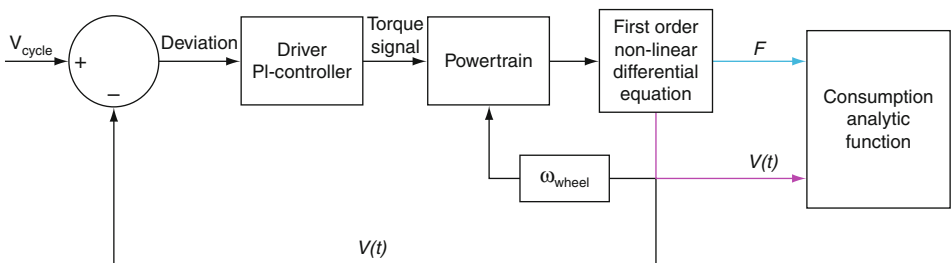
2.2 Calculation Model

The Matlab Simulink-based program, developed at the Vienna University of Technology, calculates the energy consumption of any vehicle for typical operating conditions. This longitudinal vehicle dynamics simulation model, represented by the flow chart in ► Fig. 12.4, includes powertrain submodels according to the individual propulsion concepts (ICE, hybrid, fuel cell, and BEV) addressed in this publication.

The program accepts detailed route profiles from GPS tracking as input data. Depending on the initial setup of the integrated “driver,” a closed loop traction force controller, the program calculates its own driving parameters, like powertrain torque, vehicle speed, and acceleration. The final outputs of the simulation are distance related energy and fuel consumption, and CO₂ emissions of the car are provided in g/s with a time base of 100 ms.

As inputs, the following parameters must be entered:

- Technical data of the vehicle, such as gross weight, aerodynamic factors, coefficient of rolling resistance, wheel dimensions, inertia of wheels, and differential gears



■ Fig. 12.4

Flow chart of the simulation program

- The selected reference driving cycle
- The driver behavior by parameterizing the closed loop traction force controller
- Parameters of the analytic powertrain fuel consumption function

The simulation is based on the following principle:

- Fourth order fixed step Runge–Kutta iteration method
- Hundred milliseconds time step for the simulation

► [Figure 12.5](#) shows the top hierarchy level mask of the simulation model. In the block *driving pattern* the vehicle velocity and the slope of the simulated road is set for every simulation step. The electronically simulated *driver* is generating a signal according to the difference between targeted and actual velocity, which is used to determine the necessary acceleration or braking power. The submodel *propulsion system* converts the accelerator signal to output torque and feeds it to the final drive subsystem, which includes the *blocks differential gear, brakes, wheels, driving resistances* and *energy consumption model*. This subsystem represents the final part of the drivetrain simulation: the vehicle inertia, the wheels, and their coupling to the road. It also incorporates a brake submodel that can apply, with the appropriate input signal, a brake torque on the wheels. This torque acts on the driveline in addition to the reaction imposed by the wheel contact with the road.

Finally, the performance of the vehicle with regards to fuel and energy consumption is calculated in the associated block. It incorporates a calculation submodel based on analytic functions of fuel consumption for any vehicle speed and traction force combination.

All important physical data are available as variables. In principal, all signals and state variables in the model can be displayed and saved for further analysis. Due to this layout, the model can be used to carry out parameter studies.

The simulation model was parameterized according to a warmed up combustion engine, fuel cell stack, or high power battery. Sufficient correlation could be achieved by using data from real-world measurements of various drivetrain technologies for validation (Cachón and Pucher 2007; Sekanina et al. 2007). Using this model and representative driving pattern, the power demand and energy consumption of the simulated vehicles could be calculated in detail.

2.3 Representative Real-World Driving Routes

A characteristic cross section of vehicles has already been analyzed at real-world conditions (Pucher et al. 2008b). To ensure repeatability and as a basis for comparison, standard driving routes in a highly populated area were defined. Among them are high-use main traffic routes such as freeways, to provide a good overview of the typically occurring traffic conditions.

These real-world driving cycles as well as standardized driving cycles, like the New European Driving Cycle (NEDC), Urban Driving Cycle (UDC), and Extra Urban Driving

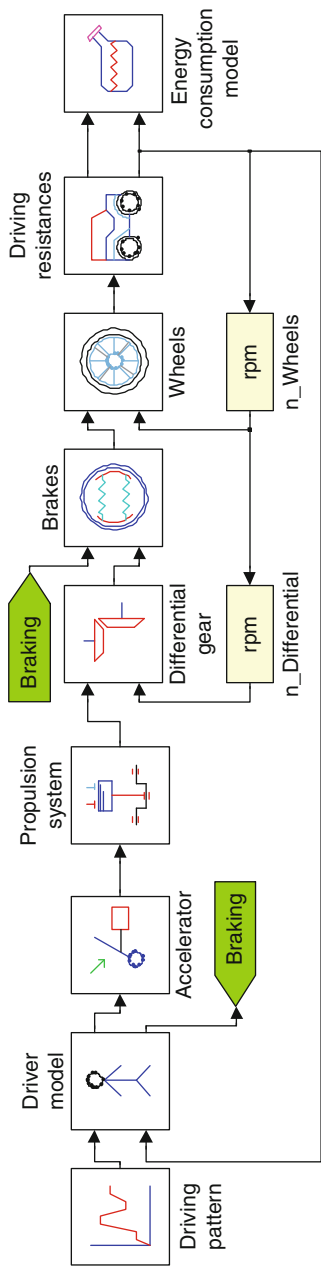


Fig. 12.5
Simulation model mask

■ Table 12.1

Standard driving routes used for simulation

Driving route	Average velocity, [km/h]	Length [km]
NEDC	33	11
UDC	19	4
EUDC	65	7
Freeway 100	97	32
Freeway 130	125	41

Cycle (EUDC) with their representative average velocity were chosen for the simulations. For details please refer to ● [Table 12.1](#). The driving cycles “Freeway 100” and “Freeway 130” correspond to test drives on the freeway where the vehicle is accelerated and kept at an average speed of 100 km/h or 130 km/h, respectively.

2.4 Power Requirement of the Vehicle

According to the vehicle dynamics, the driving resistance F_w is calculated as the sum of rolling resistance F_{Ro} , aerodynamic drag F_L , acceleration resistance F_a , gradient resistance F_{sb} , and the cornering resistance F_K (van Basshuysen and Schäfer 2004).

The power demand P_w which must be transmitted through the driven wheels to overcome the driving resistance is as follows:

$$P_w = F_w v = \left(f_r mg + \frac{1}{2} \rho c_w A v^2 + e_i m a + mg \sin \alpha + f_k mg \right) v$$

where f_r is the rolling resistance coefficient, and it is supposed to be constant at velocities below 45 m/s, m is the vehicle weight, and g is the gravitational acceleration. $\rho = p/(R^* T)$ is the air density with p as the ambient pressure, R is a gas constant, and T is the ambient temperature. c_w is the drag coefficient, and A is the maximum vehicle cross section. e_i is the rotating mass factor. The inertia values of the wheels, driveshaft, gearbox, and differential have been considered for the calculation of the rotating mass factor. $a = dv/dt$ is the vehicle acceleration. $\tan \alpha = h/l$ with h as the height of the projected distance l . f_k is the cornering coefficient.

In the formula, P_w is calculated in W, F_w in N, and v in m/s.

Since there are no unknowns in the equation, the required driving power can be computed for any given point.

Besides the different driving profiles, two scenarios with respect to real ambient conditions and power requirements of the auxiliary devices of the car were defined. The ambient scenario which represents the type approval conditions used for vehicle rating served as a basis for comparison. The real-world scenario involves average annual weather

conditions. Depending on the drivetrain technology, average annual power demand for the car auxiliaries were calculated (Shen et al. 2005). ● Table 12.2 shows the power demand of the auxiliary systems for each scenario and each vehicle concept. Conventional vehicles with IC engine, hybrid electric, and fuel cell vehicles profit from the combined

■ Table 12.2
Power demand of auxiliary systems depending on scenario

Scenario	Powertrain technology	Auxiliary systems	Power demand auxiliaries [kW]
Type approval conditions (25°C ambient temperature)	Internal combustion engine(CI-ICE)	Heating system	0
		Air conditioning	0
		Other	0.2
		Total	0.2
	Hybrid electric (SI-ICE)	Heating system	0
		Air conditioning	0
		Other	0.2
		Total	0.2
	Fuel cell electric	Heating system	0
		Air conditioning	0
		Other	0.2
		Total	0.2
	Battery electric	Heating system	0
		Air conditioning	0
		Other	0.2
		Total	0.2
Real-world scenario annual average (10°C ambient temperature)	Internal combustion engine(CI-ICE)	Heating system	0
		Air conditioning	0.5
		Other	0.5
		Total	1
	Hybrid electric (SI-ICE)	Heating system	0
		Air conditioning	0.5
		Other	0.5
		Total	1
	Fuel cell electric	Heating system	0
		Air conditioning	0.5
		Other	0.5
		Total	1
	Battery electric	Heating system	2
		Air conditioning	0.5
		Other	0.5
		Total	3

heat and power generation in IC engines and fuel cell stacks. On average, less heating power is required for these drivetrain technologies in comparison to battery electric vehicles. “Other auxiliary systems” cover basic vehicle electronics and electric devices.

3 Energy Storage Systems

The use of batteries, or more precisely rechargeable electrochemical secondary cells, for vehicles with electric powertrains has focused attention back on the problem of energy density. The following electrochemical energy converters are in use today:

Lead Acid Battery: Still the most used rechargeable battery in automotive applications. Used as a starter battery and as a traction battery.

Nickel-Metal-Hydride Battery (NiMH): Widely used in hybrid electric vehicles due to much higher energy density compared to lead acid batteries.

Lithium-Ion Battery (Li-Ion): Currently the battery with the highest energy density among all rechargeable electrochemical cells. Needs advanced charging and temperature management system.

Fuel Cell: Also supplies electric energy by converting chemically bound energy. Since the “heavy” reactant oxygen is taken from air and therefore does not need to be carried with the other reactant (e.g., hydrogen); the energy density is orders of magnitude higher.

The following table reflects the comparison of the energy density in MJ/kg and specific weight of the most common energy carriers and storage systems (● [Table 12.3](#)). The column on the right displays the weight of the particular carrier or storage necessary to contain the same energy as 1 kg of gasoline. Still, even most recent battery systems are about 150 times heavier than conventional fuels. The only energy carrier with a higher gravimetric energy density than liquid fuels is hydrogen. The data also show that 1 kg of hydrogen is equal to 400 kg of lithium-ion batteries. It has to be noted, however, that hydrogen needs to be stored in high pressure tanks or cryogenically.

■ **Table 12.3**
Energy density of storage systems

Energy carrier	Energy density [MJ/kg]	Mass of the energy carrier containing the same energy as 1 kg Gasoline [kg]
Gasoline	43	1.0
Diesel	43	1.0
Natural gas (CH ₄)	50	0.9
Hydrogen	120	0.4
Lead acid battery	0.11	400
NiMH battery	0.18	250
Li-ion battery system	0.30	150

4 Vehicle and Powertrain Concepts

4.1 Motivation for New Propulsion Systems

Developed regions have made considerable progress in the introduction of emission free vehicles. In California, the fleets of several big name car manufacturers account for some hundred fuel cell vehicles on the road which can be refueled at 25 hydrogen stations. For the grand opening of the Olympic Games 2008 in Beijing, 20 fuel cell vehicles, all developed in China, were used for the transport of VIPs.

The USA initiated the “Freedom Car Program” with the intention of independence from crude oil. The goal is the broad introduction of fuel cell vehicles between 2020 and 2025. California is presumed to be the leading region worldwide in the area of hydrogen infrastructure. An important driver of this innovation is the Californian Zero Emission Program (ZEV), which requires that a fixed percentage of cars sold in California be low or zero emission vehicles. Alternatively, automobile manufacturers are allowed to fulfill these percentages through substitution of several fuel cell vehicles (California Alternative Compliance Path, ACP). On the other hand, the US Department of Energy (DOE) is funding the development of electric vehicles with more than two billion dollars under the framework of the “American Recovery and Reinvestment Act” (ARRA). More than half of the funding is devoted to research and development of battery energy systems.

The Japanese Ministry of Economy, Trade and Industry (METI) has introduced a roadmap for batteries for electric vehicles, which includes a three-fold increase in energy density and cost reduction down to 20% of the current level within this decade. This goal is to be achieved by intense collaboration between industry, government, and universities. Japan’s “New Energy and Industrial Technology Development Organization” (NEDO) plays a key role in the realization of this battery enhancement research task. Among others, NEDO supervises the “Development of High Performance Battery Systems for Next Generation Vehicles” (Li-EAD) project which was scheduled for 2007 to 2011. Under this project, a program has been started in 2009 including 22 partners from industry and research to increase the energy density of lithium-ion batteries for electric vehicles by a factor of five compared to today’s available technology. Therefore, a joint research center has been established at Kyoto University.

The Chinese government has started a program aimed at supporting joint ventures and other cooperation between foreign and Chinese companies. Additionally, domestic businesses are being funded for the production of lithium-ion batteries of any kind. More than 40 companies have started the production of LiMn_2O_4 , LiCoO_2 , and LiFePO_4 cathodes, achieving energy densities of 130 Wh/kg and cost reductions of up to one third compared to those available on the American and European markets. Various projects with fuel cell vehicles are currently ongoing at Tonji University in Shanghai.

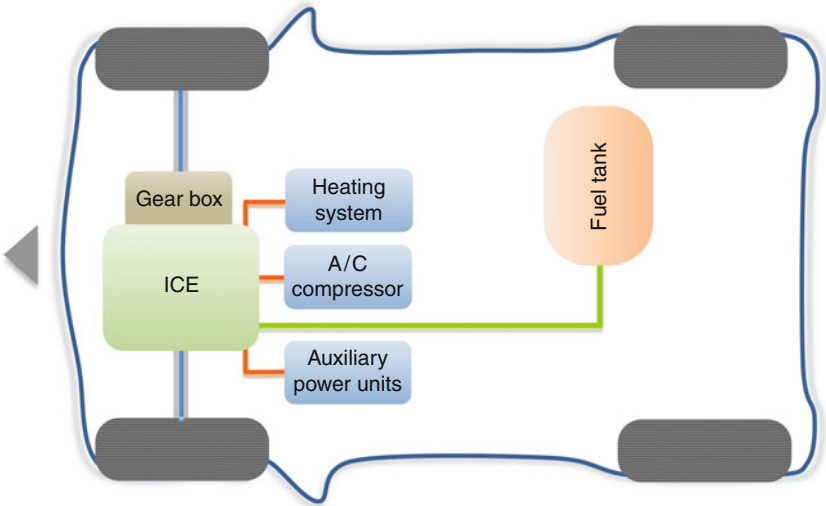
The European commission addresses electro-mobility within the framework of the Green-Cars-Initiative of the European Economic Recovery Plan. Together with industry

partners, one billion Euros should be invested in research and development for this undertaking until 2013. An ad-hoc advisory board was funded, which includes members from the European Road Transport Research Advisory Council (ERTRAC), the European Technology Platform on Smart Systems Integration (EPoSS), and SmartGrids. Together with the European Council for Automotive R&D (EUCAR) and the European Association of Automotive Suppliers (CLEPA) proposals for the arrangement of the Green-Cars-Initiative have been made and several expert workshops on batteries are being hosted together with the European Commission. A roadmap for the European industry has been created.

4.2 Internal Combustion Engine Vehicle

The vehicle used as a reference for methodical comparison of the real-world energy consumption is a statistically average European compact class vehicle with compression ignition (CI) engine and manual six speed gearbox. This car stated the interior size and the trunk of all other vehicle concepts. General physical properties of the vehicles with electric powertrain including curb weight, dimensions, and tractive resistance were derived from the reference vehicle. For vehicles with more voluminous or heavier propulsion systems like fuel cell or battery electric, a significant increase of aerodynamic resistance and weight will occur (🔗 Fig. 12.6).

The following table shows the physical properties of the reference vehicle used for the simulations (🔗 Table 12.4).



■ Fig. 12.6
Schematic layout of an internal combustion engine vehicle

■ Table 12.4
Specifications of the reference vehicle

General CI vehicle	
Curb weight	1,400 kg
Air drag coefficient, c_w	0.31
Cross sectional area, A	2.22 m ²
Rolling resistance	0.012
Tank capacity	50 L (Diesel)
CI engine	
Max. power	105 kW
Gearbox	
Type	6-Speed manual

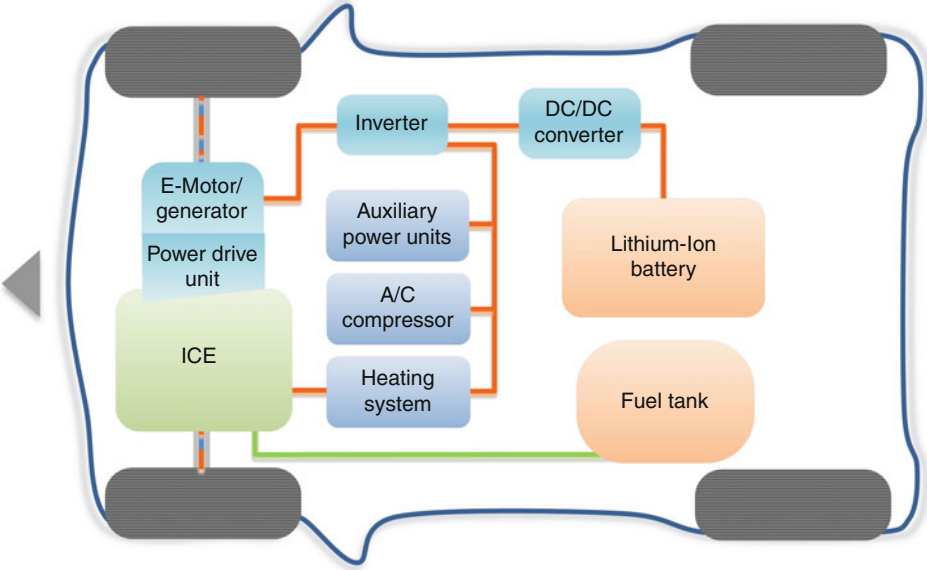
4.3 Hybrid Electric Vehicle

The drivetrain of the hybrid electric vehicle used in this comparison has a power-split structure, which is a combination of parallel and series hybrid. It incorporates a three shaft planetary gearset to distribute power from the IC engine through a mechanical and an electrical path (Liu et al. 2005). In the electrical path, power from the engine is converted in the generator and is either stored in the battery or transferred to the electric motor which, in addition to the IC engine via the mechanical path, propels the vehicle.

Usually, both paths are active for power-split hybrids even when no power is transferred to the battery. The vehicle is also able to run in electric-only mode. The overall system efficiency depends on the power distribution. While the efficiency of the mechanical path is comparable to conventional powertrains, the efficiency of the electrical path is limited due to multiple power conversions. Brake energy recovery is also possible by using the electric motor as a generator (► Fig. 12.7).

In urban driving conditions, the IC engine only runs when necessary and if it is in use then it operates at a high efficiency, due to shifting of the engine operating range to map points with best fuel efficiency. At high vehicle velocities, some of the power has to be transmitted through the electric path at decreased efficiency (Guzzella and Sciarretta 2005). In comparison to conventional ICE cars, power-split hybrid vehicles are less efficient at high velocities.

The physical properties of the hybrid electric vehicle which were used for the simulations are shown in the table below (► Table 12.5):



■ Fig. 12.7
Schematic layout of a hybrid electric vehicle

■ Table 12.5
Specifications of the hybrid electric vehicle

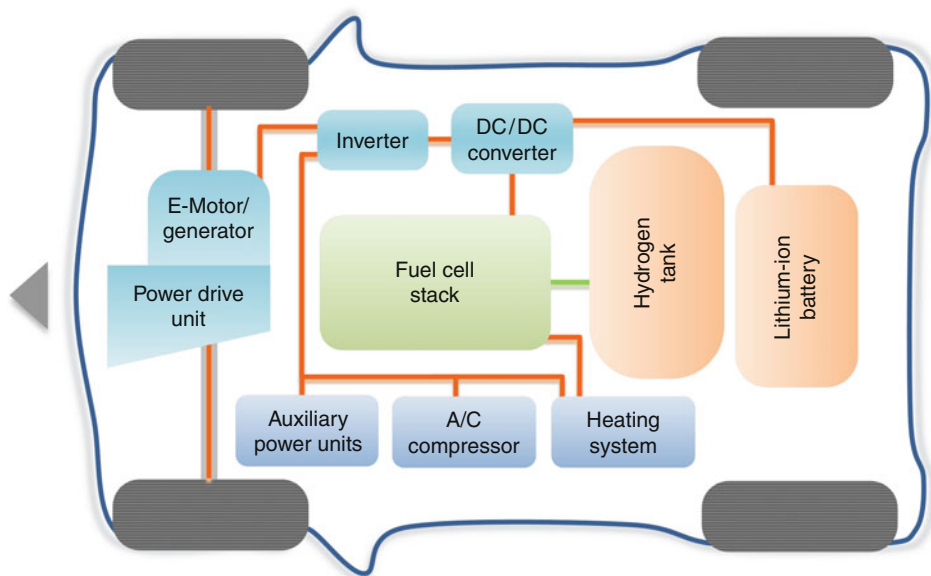
Hybrid electric vehicle	
Curb weight	1,500 kg
Max. system power	100 kW
Air drag coefficient, c_w	0.31
Cross sectional area, A	2.22 m ²
Rolling resistance	0.012
Tank capacity	45 L (Gasoline)
SI engine	
Max. power	75 kW
Electric motor/generator	
Type	Permanent magnet synchronous
Max. power output	60 kW
Battery	
Type	Nickel-metal hydride
Capacity	1.5 kWh
Weight	50 kg

4.4 Fuel Cell Electric Vehicle

Fuel cells are efficient energy converters. In contrast to IC engines, which are based on thermodynamics (heat is converted to motion), fuel cells directly convert the chemically bound energy in hydrogen to electrical energy. The products of the controlled reaction between hydrogen and oxygen are heat and water. Single fuel cells have a thickness of about 2 mm and generate a voltage of less than 1 V. Therefore, hundreds of fuel cells are combined to stacks that are able to generate the necessary voltage of more than 200 V to propel a vehicle. The reaction temperatures of commonly used proton exchange membrane fuel cells (PEM) are usually around 80°C.

A fuel cell powertrain usually consists of a hydrogen storage system, fuel cell stack, powertrain buffering battery, and electric motor (Nitsche et al. 2005). Hydrogen is stored in the light weight hydrogen pressure tanks at 700 bar and can be refilled as easily as gasoline. The heart of the fuel cell drivetrain is the stack, where the current for the electric motor is produced. The battery for the fuel cell vehicle used in this concept is a lithium-ion type with a capacity of 1.5 kWh. It is comparable to the battery size of a hybrid vehicle. The electric motor is a permanently excited synchronous motor with an output of 75 kW (► Fig. 12.8).

The following table gives an overview of the physical properties of the fuel cell electric vehicle used in this investigation (► Table 12.6):



■ Fig. 12.8

Schematic layout of a fuel cell electric vehicle

Table 12.6
Specifications of the fuel cell electric vehicle

Fuel cell electric vehicle	
Curb weight	1,500 kg
Air drag coefficient, c_w	0.32
Cross sectional area, A	2.38 m ²
Rolling resistance	0.012
Tank capacity	4 kg H ₂ @ 700 bar
Electric motor/generator	
Type	Permanent magnet synchronous
Max. power output	75 kW
Battery	
Type	Lithium-ion
Capacity	1.5 kWh
Weight	50 kg

4.5 Battery Electric Vehicle

The electric energy necessary to propel a battery electric vehicle is per definition not generated on board from another energy carrier but is transferred from an external source by the battery charger to the battery. Therefore, the charging device is the interface between the grid and the system battery-EV. The following figure shows the drivetrain structure of the battery electric vehicle concept (● Fig. 12.9).

The battery is one of the most important components of the drivetrain. Essential vehicle properties are characterized by the battery technology. The electric energy from the grid is stored electro-chemically and supplied to the electric motor. During the charging and discharging of batteries heat is created, hence some energy is lost. Therefore, the charging efficiency is defined as the quotient of extractable charge to input charge. In general, the charging efficiency is reduced significantly through fast charging and discharging. The simulations showed that the storage capacity of a standard battery electric vehicle should be at least about 24 kWh. The performance of the energy storage depends not only on the battery itself but also on the battery management system (Jeong et al. 2005). This system monitors the electrical and thermal status of the battery to optimize the operation behavior for maximum lifetime and economic efficiency. In order to maximize the lifetime of the battery, the state of charge (SOC) should be kept in a safe region between total discharge and overcharge. Therefore, the whole storage capacity is never utilized.

The inverter unit in a BEV consists of several components including the electronic control unit (ECU), the boost converter, the DC–DC converter, and the inverter itself.

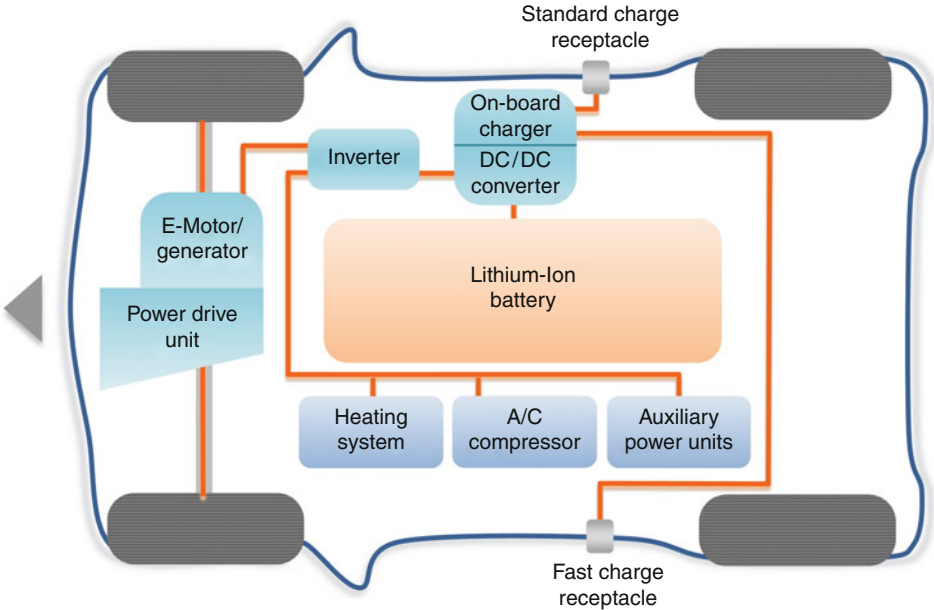


Fig. 12.9
Schematic layout of a battery electric vehicle

Table 12.7
Specifications of the battery electric vehicle

Battery electric vehicle	
Curb weight	1,700 kg
Air drag coefficient, c_w	0.32
Cross sectional area, A	2.38 m ²
Rolling resistance	0.012
Electric motor/generator	
Type	Permanent magnet synchronous
Max. power output	75 kW
Battery	
Type	Lithium-ion
Capacity	24 kWh
Weight	300 kg

The electric motor for the battery electric vehicle used in this study is a permanently excited synchronous motor with a maximum output of 75 kW. The following table shows typical specifications of the battery electric vehicle used in the simulations (Table 12.7).

5 Energy Consumption

With reference to [Sect. 4.2.4](#), a scenario which depicts the ambient conditions of the type approval test was chosen as a basis for comparison of the energy consumption. The real-world scenario representing environmental conditions averaged over a year for a region with typical continental climate.

5.1 Tank-To-Wheel Energy Consumption

The term “tank-to-wheel” refers to the energy transfer chain from the on-board energy storage system, typically fuel tank, battery, or compressed hydrogen, to the wheels during vehicle operation.

In this chapter, the final results of the simulations of the car with internal combustion engine and various vehicles with electric drivetrains are presented. The simulation results are sorted by driving conditions and propulsion concepts.

For the calculations of the type approval scenario, standard value of 25°C ambient temperature was used. The power demand of basic auxiliary and electronic systems, essential for the propulsion of a passenger car, was considered with 0.2 kW accordingly to [Table 12.2](#). All other auxiliary systems such as heating, ventilating, and air conditioning (HVAC) and comfort systems were deactivated.

The real-world scenario refers to a yearly average temperature of 10°C and includes the power demand of all auxiliary systems. As discussed previously, the conventional vehicle with CI engine, the hybrid electric, and the fuel cell vehicle are benefitting from the combined heat and power generation of the IC engines or the fuel cell stack. These vehicle concepts do not need additional heating for this scenario. In contrast, the battery electric vehicles need to devote two kW on average for heating. In [Table 12.8](#) the results of the detailed analysis are presented.

The results of the simulations show high distance related energy consumption levels for urban driving cycles compared to extra urban cycles across all vehicle concepts. For driving routes like inner city, where the average power demand for driving is less than two kW, all vehicle concepts show relatively high consumption levels, which result in low overall efficiencies.

This phenomenon is especially noticeable for the ICE vehicle. The efficiency drops to less than 20% in driving cycles with low average velocity. Here the IC engine is operating in an unfavorable area of the engine map. In contrast, high efficiencies and lower consumptions are achieved on the freeway, where the engine load is higher.

In comparison to conventional propulsion systems, hybrid vehicles show an increased efficiency in standardized driving cycles. The reference vehicle with diesel engine achieved an overall efficiency of 25% in the NEDC cycle, whereas the hybrid vehicle achieved 30%. In urban driving conditions, where the electrical path is beneficial, the hybrid drivetrain reaches 26% efficiency in contrast to 19% for the ICE powertrain. At higher velocity, the power transfer through the electrical path results in more losses than

■ Table 12.8

Comparison of the vehicle concepts energy consumption in the type approval and real-world scenarios

Driving conditions			Energy consumption [MJ/100 km]			
Test cycle	Average speed [km/h]	Ambient scenario	ICE (Diesel)	Hybrid electric (Gasoline)	Fuel cell electric	Battery electric
NEDC New European driving cycle	33	Type approval	164	136	98	71
		Real-world	188	161	108	108
UDC Urban driving cycle	19	Type approval	192	137	101	64
		Real-world	230	169	116	129
EUDC Extra urban driving cycle	65	Type approval	147	134	97	75
		Real-world	162	156	105	96
Freeway 100	97	Type approval	146	149	102	84
		Real-world	159	159	110	100
Freeway 130	125	Type approval	193	207	151	117
		Real-world	207	218	161	131

in mechanical drivetrains. The hybrid electric vehicle reaches the highest range among the vehicles with electric powertrains, which is due to the use of gasoline as on board energy storage.

The overall efficiency of the fuel cell vehicle at type approval conditions is 46%, due to high efficiencies of the fuel cell stack and the electric motor. Even in real-world conditions, the efficiency stays high. Four kilograms of hydrogen are equal to a range of 500 km in the NEDC. In urban driving conditions ranges of 400 km and on the freeway ranges of more than 300 km can be achieved.

The battery electric vehicle is able to achieve an efficiency of about 70% in type approval scenarios, a much higher value compared to levels achievable by conventional vehicles in standard driving cycles. In the real-world scenario, the tank-to-wheel efficiency drops down to 35% in urban driving and to 60% at higher velocities. These values are strongly dependent on the efficiency of the electric motor and the discharge efficiency of the battery. Additionally, the range of a battery electric vehicle is very limited due to the battery capacity. In the NEDC, the battery electric vehicle is able to achieve a range of more than 120 km. During real-world driving the maximum range is reduced to 70 km.

5.2 Primary Energy Consumption

After focusing on the real-world, tank-to-wheel energy consumption of passenger cars is necessary for a consistent comparison to look into where this energy is coming from. Therefore, calculations based on a typical electric energy mix for a region with 40% non-carbon primary sources were performed. The efficiencies of the different stages of the energy supply chain from primary energy to tank are presented in [Table 12.9](#). These values influence significantly the total consumption of primary energy carriers.

[Table 12.10](#) shows the summary of most important results of all examined vehicle concepts in the NEDC driving cycle. For every propulsion concept, the two NEDC lines are distinguished by the subscript “type approval” and “real-world” which defines the ambient scenarios.

The consumption is stated, depending on the propulsion concept, in liter diesel or gasoline per 100 km for the ICE and the hybrid vehicle, in kilograms H_2 per 100 km for the fuel cell vehicle, and in kilowatt-hours per 100 km for the battery electric vehicle. For a general comparison, the energy consumption is shown in MJ/100 km as well for “tank-to-wheel” and “primary energy” approach. The “maximum range” was calculated from the energy consumption and the tank size or battery capacity.

The “primary energy consumption” column shows substantially higher results in comparison to the tank-to-wheel calculations especially for battery electric vehicles, taking into account the entire energy supply chain.

■ Table 12.9

Efficiencies of energy generation process stages (primary energy to tank)

Energy source	Process stages	Efficiency [%]
Fossil fuels	Production and distribution	90
Electricity	Storage of electricity in accumulator batteries (regular charge)	75
	Average power generation efficiency	50
Hydrogen (from methane reformation)	Production of H_2 from methane reformation	75
	Compression of H_2 for pressure fueling at 700 bar	85
Hydrogen (from electrolysis)	Average power generation efficiency	50
	Production of H_2 from electrolysis	75
	Compression of H_2 for pressure fueling at 700 bar	85

■ Table 12.10
Summarized comparison of tank-to-wheel and primary energy consumption of the vehicle concepts

Propulsion concept	Consumption	Tank-to-wheel energy consumption [MJ/100 km]	Primary energy consumption [MJ/100 km]		Maximum range [km]
ICE diesel					
	Liter _{Diesel} /100 km				
NEDC Type approval	4.6	164	182		1,070
NEDC Real-world	5.3	188	208		940
Hybrid electric					
	Liter _{Gas} /100 km				
NEDC Type approval	4.2	136	151		1,080
NEDC Real-world	4.9	161	179		910
Fuel cell electric					
	kg H ₂ /100 km		Methane reformation	Electrolysis	
NEDC Type approval	0.8	98	153	306	490
NEDC Real-world	0.9	108	169	338	445
Battery electric					
	kWh/100 km				
NEDC Type approval	20	71	189		122
NEDC Real-world	30	108	289		80

6 Summary

The Californian zero emission regulations and the European CO₂ limits are the main driving forces toward zero emission propulsion systems. Due to this situation, the main focus of this chapter was a methodical comparison of a statistically average car with a new electric powertrain such as hybrid, fuel cell, and battery electric relative to the conventional combustion engine.

The comparative analysis showed that for typical usage in urban, extra urban, and freeway real-world traffic the hybrid system and, with some range limitations, the fuel cell system can provide a similar usability compared to regular systems with internal combustion engines. Battery electric vehicles can only be used for short distances due to the limited range in real-world conditions.

Based on a primary energy calculation in general, the energy consumption of the hybrid and the fuel cell vehicle is better than the consumption of the other propulsion concepts that were examined. Battery electric vehicles have higher energy consumption

owing to inefficiencies in power production and transmission as well as charging and discharging losses.

It must also be stated that cars with fuel cell and battery electric propulsion systems have the considerable benefit of local emission free driving.

7 Symbols and Abbreviations

Symbol	Unit	Explanation
a	m/s^2	Acceleration
A	m^2	Maximum vehicle's cross section
CO_2	–	Carbon dioxide
CO	–	Carbon monoxide
c_w	–	Drag coefficient
e_i	–	Rotating mass factor
f_r	–	Rolling resistance coefficient
f_k	–	Cornering coefficient
F_a	N	Acceleration resistance
F_K	N	Cornering resistance
F_L	N	Aerodynamic drag
F_{Ro}	N	Rolling resistance
F_{st}	N	Gradient resistance
F_w	N	Driving resistance
g	9.81 m/s^2	Gravitational acceleration
m	kg	Curb weight
NO_x	–	Nitrogen oxides
p	mm Hg	Ambient pressure
P_W	W (kW)	Power demand
R	–	Gas constant
t	s	Time
T	k	Ambient temperature
v	m/s	Velocity
ρ	kg/m^3	Air density

Abbreviation	Explanation
ACP	Alternative Compliance Path
DC/DC	Direct Current/Direct Current
ARRA	American Recovery and Reinvestment Act

Abbreviation	Explanation
BEV	Battery Electric Vehicle
BMS	Battery Management System
CI	Compression Ignition
CLEPA	European Association of Automotive Suppliers
DOE	US Department of Energy
ECU	Electronic Control Unit
EPoSS	European Technology Platform on Smart Systems Integration
ERTRAC	European Road Transport Research Advisory Council
EUCAR	European Council for Automotive R&D
EUDC	Extra Urban Driving Cycle
GPS	Global Positioning System
HC	Hydrocarbons
H ₂	Hydrogen
HVAC	Heating, Ventilating And Air conditioning
ICE	Internal Combustion Engine
METI	Japanese Ministry of Economy, Trade and Industry
NDIR	Non-Dispersive Infrared Absorption
NEDC	New European Driving Cycle
NEDO	Japan's New Energy and Industrial Technology Development Organization
NiMH	Nickel-Metal-Hydride Battery
PEMS	Portable Emission Measurement System
SOC	State of Charge
UDC	Urban Driving Cycle
ZEV	Zero Emission Program

Acknowledgments

We would like to thank in particular Dr. James Girard for reviewing this contribution.

References

Cachón L, Pucher E (2007) Fuel consumption simulation model of a CNG vehicle based on real-world emission measurement. SAE Technical Paper Series, Paper-Nr. 2007-24-0114

Guzzella L, Sciarretta A (2005) Vehicle propulsion systems, introduction to modeling and optimization. Springer, Berlin/Heidelberg. ISBN 3-540-25195-2

Jeong KS, Lee WY, Kim CS (2005) Energy management strategies of a fuel cell/battery hybrid system using fuzzy logics. J Power Sour 145:319–326. doi:10.1016/j.jpowsour.2005.01.076, Elsevier 2005

Liu J, Peng H, Filipi Z (2005) Modeling and analysis of the Toyota hybrid system. In: Proceedings of the IEEE/ASME international conference on

- advanced intelligent mechatronics, Monterey, California. IEEE-Paper 0-7803-9046-6
- Nitsche CH, Weiss W, Fosmoe R, Schroedl S, Pucher E (2005) Predictive maintenance and diagnostic concept for a worldwide fuel cell vehicle fleet. In: Proceedings of the 21st Worldwide battery, hybrid and fuel cell electric vehicel Symposium, Monte. Paper-Nr. FFP298, 16S
- Paces M (2007) Simulation der streckenbezogenen CO₂-Emissionen eines Hybridfahrzeugs auf Basis von Real-World Messungen, Vienna University of Technology
- Pucher E, Tóth D, Cachón L, Weissenberger D (2008a) Real World Emission measurement of conventional and hybrid light duty vehicles with increased bio fuel blends. In: FISITA 2008 World Automotive Congress, Munich. F2008-09-039
- Pucher E, Tóth D, Cachón L, Raetzsch T, Weissenberger D (2008b) Real World and chassis dynamometer emission measurement of a turbocharged gasoline vehicle with increased bio fuel blend. SAE Technical Paper Series, Paper-Nr. 2008-0-1768, Shanghai
- Sekanina A, Pucher E, Gruber K, Kronberger H (2007) Fuel Cell vehicle efficiency optimization by advanced fuel cell design and drive train simulation, SAE Technical Paper Series, Paper-Nr. 2007-24-0072
- Shen ZJ, Chen X, Masrur A, Garg VK, Soltis A (2005) Optimal power management and distribution in automotive systems. In: Emadi A (ed) Handbook of automotive power electronics and motor drives. Taylor and Francis, Baco Raton, Chapter 5
- Simic D, Pucher E (2009) Saving potential of HDV auxiliaries energy consumption determined by entire vehicle simulation. In: Proceedings of the vehicle power and propulsion conference, Lille, VPPC '09. IEEE, 2009, ISBN: 978-1-4244-2600-3; S. 785–S. 789
- van Basshuysen R, Schäfer F (2004) Internal Combustion Engine, SAE International, ISBN 0-7680-1139-6

Section 4

Positioning, Navigation, and Trajectory Control

Peter Handel

13 Global Navigation Satellite Systems: An Enabler for In-Vehicle Navigation

John-Olof Nilsson · Dave Zachariah · Isaac Skog
School of Electrical Engineering, KTH Royal Institute
of Technology, Stockholm, Sweden

1	<i>The GNSS Technology</i>	313
1.1	GNSS for Vehicle Positioning	314
1.2	GNSS Technology Limitations	314
1.3	Current GNSSs	315
2	<i>Principles</i>	316
2.1	The Geometry of Satellite Positioning	316
2.2	Signal Properties	318
2.3	System Components and Structure	320
3	<i>Theory</i>	322
3.1	Pseudorange and Position Relation	322
3.1.1	Position Estimation	322
3.2	Received Signal and Pseudorange Relation	323
3.2.1	Undistorted Signal	324
3.2.2	Random Distortion	324
3.2.3	Systematic Distortion	325
4	<i>Practice</i>	327
4.1	Position Estimation	327
4.1.1	LS Position Estimation	327
4.1.2	MMSE Position Estimation	329
4.2	Measuring Pseudoranges	329
4.2.1	Signal Acquisition	330
4.2.2	Carrier and Code Phase Tracking	331
4.2.3	Carrier Phase Tracking	332
4.2.4	Navigation Data	333
4.3	Pseudorange Error Sources	333

4.4	GNSS Receivers	335
4.4.1	GNSS Antennas	335
4.4.2	RF Front End and ADC	336
4.4.3	Hardware Signal Acquisition Implementation	336
4.4.4	Hardware Pseudorange Tracking Loop Implementation	337
4.4.5	Position Estimation	337
4.4.6	Software Receivers	337
5	<i>Differential GNSS</i>	337
5.1	Principles	338
5.2	Augmentation Systems	338
5.2.1	Satellite-Based Augmentation Systems	339
5.2.2	Ground-Based Augmentation Systems	340
6	<i>Conclusion and Further Reading</i>	340

Abstract: The emergence of global navigation satellite systems (GNSSs) has enabled tremendous development in vehicular navigation for various applications. The GNSS technology provides a unique global positioning capability with meter-level accuracy at a low hardware cost and zero marginal infrastructure cost. The GNSSs work by using the satellites as radio beacons, broadcasting a satellite-specific signal and their own position. The range to the satellites is measured up to a common clock offset, and any user equipped with a GNSS receiver capable of receiving the signal from four or more satellites can position itself by multilateration. The position, being a fundamental piece of information for automatizing and facilitating location-dependent system interaction and services, makes the GNSS an enabling technology for many intelligent vehicle and transportation system capabilities.

This chapter will focus on introducing the basic principles of the GNSS technology and the signal processing that allows the GNSS receiver to determine its position: in 🔍 Sect. 1, the technology, its limitations, and currently available GNSSs are reviewed; in 🔍 Sect. 2, the principles of the GNSS positioning, the signal characteristics, and fundamental components are discussed; in 🔍 Sect. 3, the theoretical relations governing the positioning are presented; in 🔍 Sect. 4, implementation-related issues, error sources, and GNSS receivers are discussed; in 🔍 Sect. 5, the method of differential GNSS is introduced and current augmentation systems are reviewed; and finally in 🔍 Sect. 6, conclusions are drawn and references, for further reading about different aspects of the GNSS technology, are given.

1 The GNSS Technology

The global navigation satellite system (GNSS) technology is a radio positioning technology with global coverage based on satellite infrastructure. The most well-known example of an existing system is probably the US NAVSTAR Global Positioning System (GPS) but other systems exist. The systems have a typical positioning accuracy of down to a few meters but with capability, in combination with augmentation system, of down to centimeter-level accuracy.

The GNSS technology is scalable in the sense that it allows for an unlimited number of users independent of the satellite infrastructure. The satellites work as radio beacons, broadcasting a satellite-specific signal and their own position. Any user equipped with a GNSS receiver capable of receiving the signal from four or more satellites can position itself. No information transfer takes place from the user to the satellites.

A GNSS consists typically of more than two dozen satellites orbiting the Earth in a constellation such that four or more satellites are in view from any point on Earth at any time. The satellites are controlled and monitored by a network of control and monitoring stations. The satellites transmit signals in frequency bands $\sim 2\text{--}40$ [MHz] wide in the spectral range $1.2\text{--}1.6$ [GHz]. The signal transmissions from the satellites are synchronized, and a user equipped with a GNSS receiver positions itself by multilateration based on time (differences) of arrival.

1.1 GNSS for Vehicle Positioning

For vehicles and transportation systems, the GNSS technology can provide a unique global position capability for vehicles as well as personnel, freight, and assets. The position, being a fundamental piece of information for automatizing and facilitating location-dependent system interaction and services, makes the GNSS an enabling technology for many intelligent vehicle and transportation system capabilities. Already today GNSS technology helps many drivers to find their way around on the road network, it is an integral part of many fleet management systems, and an increasing share of vehicles are equipped with GNSS receivers. The basic technology is mature but with a large developmental potential in new satellite signals, signal processing, low-cost receiver design, and integration with other systems. Receivers are available off the shelf with a wide range of position accuracy, flexibility, and cost. Basic receiver chips are readily available for less than €5, and the marginal cost of positioning infrastructure is zero. In summary, GNSSs are likely to provide an important information source for vehicle positioning for a foreseeable future.

The positioning capability requires vehicles to be equipped with GNSS receivers. Also, for the position information to be useful, as a minimum it requires the vehicle to carry other location-dependent information, that is, maps in a general sense. With such information, the vehicle can perform route planning and guidance. However, for the position information to be useful, apart from being a driver navigation tool, it also requires the vehicles and transportation infrastructure to have the capabilities to forward position and intentions or to request, retrieve, and store position-related information from agents with geographically overlapping interests, that is, vehicles need to communicate with each other, to the local transportation infrastructure and to the global information infrastructure. Further, in combination with other sensors such as radars, cameras, inertial sensors, and wheel encoders, it has the capability to provide sufficient information for semi- or fully autonomous vehicles.

1.2 GNSS Technology Limitations

Despite providing global positioning with an unprecedented accuracy and availability, the GNSS technology has limitations that one needs to be aware of. First of all, even though it provides a position estimate with good accuracy on a global scale, on a local scale its accuracy is typically not sufficient to determine, for example, lane position for vehicle applications. Even though the GNSS satellite signals, in combination with augmentation systems, are capable of giving centimeter accuracy under favorable conditions, achieving such accuracy with a GNSS stand-alone system, while the receiver is in motion in an environment with frequent signal path blockage and reflections – typical for vehicular applications – is difficult at best. Second, the accuracy of the position estimate is difficult to assess. That is, the integrity of the position often limits the usefulness of the mean position accuracy. Third, in urban canyons, tunnels, and other radio shadowed areas, the accuracy and integrity of the system is greatly reduced or a position might not be

available at all. In other words, the coverage of the system is not complete. Finally, the system can malfunction or be locally jammed, unintentionally or by malicious actors. In summary, the accuracy, integrity, and availability of the GNSSs as stand-alone systems are not sufficient for some demanding applications, and due to coverage and reliability, the GNSS technology cannot be solely depended upon for safety critical applications.

1.3 Current GNSSs

Currently, there are two fully operational GNSSs. The first and most widely known system is the US NAVSTAR Global Positioning System (GPS). The second GNSS with full operational capacity is the Russian Global Navigation Satellite System (GLONASS). Two other systems, the European Galileo and the Chinese COMPASS systems, are as of 2011 under construction. Also some regional systems, based on similar technology but without global coverage intentions and satellites in geostationary orbit, exist such as the Chinese BeiDou Navigation System (predecessor of the COMPASS system), or is under construction such as the Indian Regional Navigational Satellite System (IRNSS). Summary and some technical data of the GNSS are found in the [Table 13.1](#) Cf. (Kaplan and Hegarty 2006; Gao et al. 2007; Grelief et al. 2007; Grewal et al. 2007).

The different satellite systems broadcast multiple signals centered at different frequencies. For instance, GPS satellites transmit two spread-spectrum signals centered at 1.228 and 1.575 [GHz], denoted L1 and L2, respectively. Civilian receivers use a coarse acquisition (C/A) code modulated on the L1-signal for positioning, while the US military uses a longer precision (P) code, encrypted and modulated on both the L1- and L2-signals, that allows for higher resolution and therefore more accurate position estimates.

GLONASS transmits signals according to a different principle, where each satellite signal is allocated its own frequency band but all share the same code. The satellites send two signals, L1 and L2, whose carrier frequencies are located in the intervals 1.598–1.604 and 1.243–1.249 [GHz], respectively. Like GPS, the L2-signal is reserved for a precision code used for military applications.

■ Table 13.1

Summary and technical data of the GNSS currently deployed or under construction

GNSS	Operating countries	Orbital height [km]	Orbit period [h]	Number of satellites	Frequencies [GHz]
GPS	USA	26,560	12.0	>24	1.228, 1.575
GLONASS	Russia	25,510	11.3	24	1.597–1.617, 1.240–1.260
Galileo	EU	23,222	14.1	30 planned	1.164–1.300, 1.559–1.592
COMPASS	China	21,150	12.6	35 planned	1.207–1.269, 1.561–1.590

Galileo is designed to provide three different civilian services: open, commercial, and safety of life. The code immediately used for positioning is modulated on the L1-signal at 1.575 [GHz], but the open service also includes the so-called E5-signal at 1.192 [GHz] which can be used for improved resolution. The commercial service uses an additional signal, denoted E6, located at 1.278 [GHz], for high-accuracy applications.

While the GPS reached full capabilities for civilian use in the 1990s, efforts have been made to update the system. Recent modifications include the addition of the L5-signal centered at 1.176 [GHz] for civilian use as well as modulating more civilian and military codes on the L1- and L2-signals. Modernization of the GLONASS is also underway including the introduction of code division methods similar to GPS and Galileo.

2 Principles

Positioning based on the GNSS technology is possible due to geometrical relations between the receiver and the satellites, known satellite positions, and properties of the transmitted signals. In this section, the principles of these relations and properties are described and illustrated. The required GNSS components and structure are also discussed.


2.1 The Geometry of Satellite Positioning


Consider a GNSS satellite i and a receiver arbitrarily located in space, and suppose the following data is known: the satellite position $[x_i, y_i, z_i]$ in Cartesian coordinates, the nominal speed, c , of light in vacuum, at which a given signal propagates from the satellite and its traveling time Δt_i from the satellite i to the receiver. From this the geometrical range γ_i between the receiver and the satellite i can be computed,

$$\gamma_i = c\Delta t_i \quad (13.1)$$

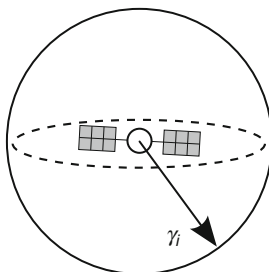
The geometrical range between the satellite and the receiver can also be expressed as a function of the unknown receiver position $[x, y, z]$,

$$\gamma_i(x, y, z) = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} \quad (13.2)$$

This relation will constrain the receiver position to lie on a sphere centered at the satellite and with a radius γ_i . The constraint is illustrated in  Fig. 13.1.

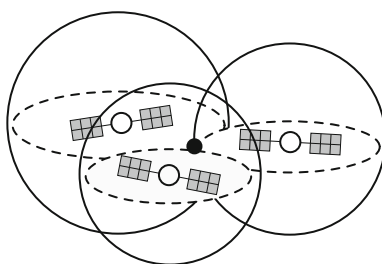
With two satellites, corresponding positions and signal traveling times (ranges), the position is constrained to the intersection of two spheres, that is, a two-dimensional circle. With three satellites, the position can in principle be uniquely determined as the intersection of three spheres illustrated in  Fig. 13.2.

The transmitted signals from each satellite repeat periodically in accordance with a satellite system clock. If the receiver follows the system clock it knows when the signals were transmitted. Then if it records when the signals are received the traveling times Δt_i from each satellite i , and therefore also the ranges γ_i , become known and it could determine



■ Fig. 13.1

Given the geometrical range γ_i from the satellite i to the receiver, the position of the receiver is constrained to a sphere centered around the satellite position $[x_i, y_i, z_i]$ and with radius γ_i



■ Fig. 13.2

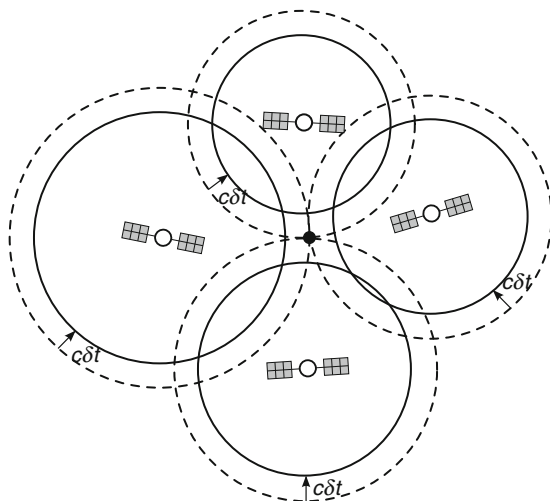
With signal traveling times (ranges) between the receiver and three or more satellites at known locations, the position of the receiver is uniquely determined by the intersection of the constraining spheres of the individual range measurements

its position. Unfortunately, one major obstacle to satellite positioning, as outlined above, is clock synchronization: The satellite and receiver clocks are not perfectly synchronized to the system clock. For instance an error of 1 [ms] in measured traveling time will lead to a range error of 3×10^5 [m], clearly inadequate for global positioning.

The difference between transmit t_i^s and receive time t_i^r of the satellite signal recorded at the receiver is thus the traveling time plus an error due to an unknown clock offset between satellite and receiver. Typically, in comparison with the system clock, the receiver clock offset is much larger than the satellite clock offset and therefore for now we assume that the satellites are perfectly synchronized with the system clock. Hence, what can be calculated based on signal measurements are not the geometrical ranges $\gamma_i = c\Delta t_i$ but the so-called *pseudoranges* $\rho_i = c(t_i^r - t_i^s)$. Let the deviation from the system clock be denoted as δt , then the relation between pseudoranges and geometric ranges are

$$\rho_i = \gamma_i(x, y, z) + c\delta t \quad (13.3)$$

Note that δt is common to all pseudorange measurements at one time instant. Combined with (► Eq. 13.2) this gives four unknowns, $[x, y, z]$ and δt , but with three



■ Fig. 13.3

Due to the additional ambiguity introduced by the receiver clock offset δt , the position can only be resolved from the pseudoranges with a minimum of four satellites

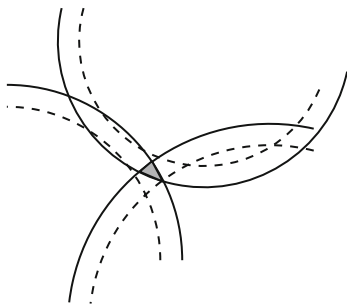
satellites, as of Fig. 13.2, there are only three constraints $i = 1, 2, 3$ available. To resolve the ambiguity, a fourth satellite equation is required. Therefore, at least four satellites must be visible for positioning. This can be thought of as the situation illustrated in Fig. 13.3 where the spheres have radii with a common unknown bias $c\delta t$ that has to be solved for. This is the geometry used in GNSS positioning.

A final small obstacle to overcome in the geometry of the positioning is the host of error sources apart from the clock offset, appearing in the transmission, propagation, and reception of the radio signal, introducing errors in t_i^r and t_i^s , and consequently also in ρ_i . In loose terms, the constraining sphere of Fig. 13.2 is really a spherical shell. In turn, the intersection of three or more spheres as of Fig. 13.2 becomes intersection of spherical shells which results in an uncertainty volume of the receiver position rather than a single unambiguous point. The intersection and uncertainty volume are illustrated in Fig. 13.4. The situation generalizes to the geometry of Fig. 13.3 and in the end the position has to be determined based on some error minimization criterion.

2.2 Signal Properties

From the description of the position principles above, the transmitted signals $s_i(t)$ from each of the mutually synchronized satellites must carry a minimum of information to the receiver:

1. A uniquely identifiable signature
2. A periodic component following the system clock with an unambiguous phase



■ Fig. 13.4

The intersection of three spherical shells resulting in an uncertainty volume of the receiver position rather than a single unambiguous point as of Fig. 13.2

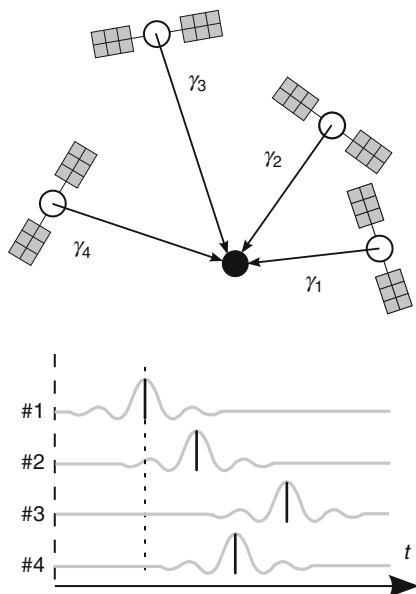
The signal signature determines the satellite i and its phase gives the corresponding pseudorange ρ_i . Since the signals are transmitted simultaneously they need to be transmitted by some multiple access method. Typically, code division multiple access (CDMA) is used. GLONASS uses frequency division multiple access (FDMA) for some of its signals. For brevity, only CDMA will be used hence forth (Kaplan and Hegarty 2006; Proakis and Salehi 2008). The relations between the geometrical ranges and the phase of the different signals are illustrated in Fig. 13.5. The current positions of the satellites could in principle be transmitted via alternative channels but are typically also encoded in a subset of the satellite signals.

The required information is embodied in an electromagnetic wave with radio frequency that is capable of penetrating the atmosphere and reaching the surface of the Earth. The GNSS satellite signals are transmitted in frequency bands $\sim 2\text{--}40$ [MHz] wide in the spectral range $1.2\text{--}1.6$ [GHz]. The transmitted power must be large enough to mitigate interference from other sources of radio waves, but still meet the constraints of the satellites. In general, since the information transfer in the signals is very small (only phase) and the signals are transmitted with spread-spectrum techniques, the transmitted power can be designed such that the signal received on Earth is very weak. The power of the received signals is typically below -153 [dBW], which is below the thermal noise floor of typical operational temperatures.

A general model for the signal is a carrier wave $\sqrt{P} \cos(2\pi f_c t)$, with frequency f_c and power P , modulated with a bipolar information signal $C_i(t)$ giving the unique signature,

$$S_i(t) = \sqrt{P} \cdot C_i(t) \cos(2\pi f_c t) \quad (13.4)$$

In the processing of the signal, we will in general assume that the information signal $C_i(t)$ is known. A discussion about the situation when the signal is not completely known is given in Sect. 4.2.4. The information signal $C_i(t)$, carrier $\cos(2\pi f_c t)$, and modulation are illustrated in Fig. 13.6. The information signals are designed as so-called pseudo-random noise (PRN) codes. This gives the signals a distinct phase.



■ Fig. 13.5

Relation between geometrical range and the phase of different signals. The signals are transmitted synchronously. The bars denote the phase of the different signals. The further the geometrical distance, the more the phase of the signal is shifted. Note that the signals only illustrate the phase shift characteristics and do not carry satellite signatures and do not resemble typical GNSS signals

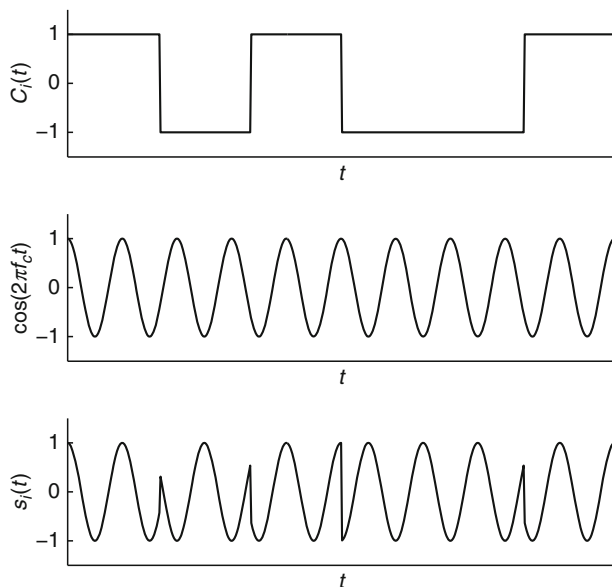
2.3 System Components and Structure

A number of fundamental components are needed in a GNSS. In general, one speaks about the space segment, the user segment, and the control segment. The required components and information transfer within the systems are illustrated in ► Fig. 13.7.

The space segment is the constellation of satellites. The satellites need to be capable of transmitting signals with properties as described in ► Sect. 2.2. The transmission needs to be synchronous throughout the system, and consequently the satellites have to carry very accurate local clocks (atomic clocks). Further, the satellites are required not only to transmit the signals described above, they must also be positioned in space so as to cover a sufficiently large area of the Earth. While local satellite positioning systems are conceivable, a global system requires that a minimum of four satellites are visible at any point on the globe at any time.

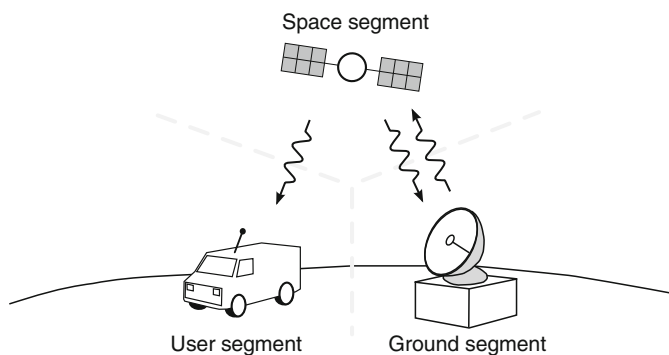
The user segment consists of the receivers. A receiver must be able to receive signals from four or more satellites to extract the signal properties as of ► Sect. 2.2, and to estimate position based on these extracted properties.

The control segment is the system controlling the satellites. Since the satellites do not constitute an autonomous system, they must engage in two-way communication with the



■ Fig. 13.6

The information signal $C_i(t)$ (top) is modulated on the carrier $\cos(2\pi f_c t)$ (middle) resulting in the transmitted signal $s_i(t)$ (bottom)



■ Fig. 13.7

The components of GNSSs can be divided into a space segments, consisting of the satellite constellation, a user segment, consisting of the receivers, and a ground segment, consisting of the control and monitoring stations

control segment which has to perform some crucial tasks for the positioning system to work: tracking the satellites, controlling their orbits, monitoring their health and the overall integrity, and maintaining the stability and precision of the high-resolution system clock. Apart from giving direct control commands, the control segment must also transfer the tracked orbit (position) of the satellites to them such that it can be relayed to the user segment.

3 Theory

Conceptually, the method of determining position from pseudoranges by multilateration in four or more dimensions is explained in [Sect. 2.1](#). However, from the principles it is not clear what the relations between position, received signal, and measured pseudoranges are. In this section, we will state the functional dependence of the measured pseudoranges on position and give the minimization criterion which will in theory determine our position estimate given pseudorange measurements. Further, we will also give the mathematical relations from which the pseudoranges are extracted from raw radio measurements using basic properties of the signals.

3.1 Pseudorange and Position Relation

The measured pseudorange $\tilde{\rho}_i$ corresponding to satellite i with position $[x_i, y_i, z_i]$ relates to the ideal pseudorange ρ_i and the position of the receiver $[x, y, z]$ via the relation

$$\begin{aligned}\tilde{\rho}_i &= \rho_i + e_i \\ &= \gamma_i(x, y, z) + c\delta t + e_i\end{aligned}\quad (13.5)$$

where

$$\gamma_i(x, y, z) = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2}$$

is the geometrical range to the satellite i , c is the speed of light, δt is the unknown receiver clock offset, and e_i is an error term accounting for nonideal signal transmission, propagation, and reception. Though not explicitly stated, the error term will have components dependent on both time, receiver position, satellite position, as well as stochastic components. We will return to this term in [Sect. 4.3](#). For now, we assume that it can be *partly* modeled and compensated for as e_i^{model} .

3.1.1 Position Estimation

Based on multiple pseudorange measurements $\tilde{\rho}_i : i \in \{1, \dots, N\}$, the receiver will have to estimate its position $[x, y, z]$. If four or more measurements are available, there will be equal or less unknowns $[x, y, z, \delta t]$ than constraints,

$$\begin{aligned}\tilde{\rho}_1 &= \gamma_1(x, y, z) + c\delta t + e_1 \\ \tilde{\rho}_2 &= \gamma_2(x, y, z) + c\delta t + e_2 \\ &\vdots \\ \tilde{\rho}_N &= \gamma_N(x, y, z) + c\delta t + e_N\end{aligned}\quad (13.6)$$

where $N \geq 4$. Unfortunately, due to imperfect models and stochastic components of the e_i -terms, the different constraints will not be consistent and a position $[x, y, z]$ cannot be solved for. Consequently, the position will have to be estimated based on some inconsistency minimization criterion. In combination with the nonlinear nature of (Eq. 13.6), this gives rise to a nonlinear minimization problem,

$$[\hat{x}, \hat{y}, \hat{z}] = \arg \min_{[x, y, z]} \left(\min_{\delta t} F(\mathbf{x}) \right) \quad (13.7)$$

where $[\hat{x}, \hat{y}, \hat{z}]$ is the position estimate, the unknown variables are collected in $\mathbf{x} = [x, y, z, \delta t]$, and $F(\cdot)$ is the criterion function. If a priori statistical information of \mathbf{x} or statistical information about the measurement errors are unavailable or ignored, the sum of the squared measurement residuals,

$$r_i(\mathbf{x}) = \tilde{\rho} - \gamma_i(x, y, z) - c\delta t - e_i^{\text{model}} \quad (13.8)$$

is commonly used as a criterion function,

$$[\hat{x}, \hat{y}, \hat{z}] = \arg \min_{[x, y, z]} \left(\min_{\delta t} \sum_{i=1}^N r_i^2(\mathbf{x}) \right) \quad (13.9)$$

lending the minimization to a least-squares (LS) problem (Kay 1998). Note that with respect to the minimization, any common component in the error terms e_i over all pseudorange measurements will be equivalent to a receiver clock offset. Hence, common mode errors in the pseudoranges will not affect the position estimate.

Alternatively, if prior statistical information of the true variables \mathbf{x}^* is known, for example, using a dynamic motion model of the receiver and a statistical measurement model of (Eq. 13.5), the mean square error (MSE) can be used as criterion function,

$$[\hat{x}, \hat{y}, \hat{z}] = \arg \min_{[x, y, z]} \left(\min_{\delta t} E(\|\mathbf{x}^* - \mathbf{x}\|_2^2) \right) \quad (13.10)$$

where the expectation is conditioned on all past pseudorange measurements, lending the minimization to a minimum mean square error (MMSE) problem (Kay 1998; Kailath et al. 2000), solved in a Bayesian estimation framework. Solutions to both these approaches, (Eq. 13.9) and (Eq. 13.10), are further discussed in Sect. 4.1.

3.2 Received Signal and Pseudorange Relation

To estimate the position, the pseudoranges $\tilde{\rho}_i$ need to be measured from the received signal. This is possible from the minimum signal properties listed in Sect. 2.2. In this section, we will look at the mathematical relations that the receiver is approximating when extracting the pseudoranges from raw radio measurements.

3.2.1 Undistorted Signal

Assume ideal signal transmission, propagation, and reception, that is, the signal components $s_i(t)$ from the satellites $i = 1, 2, \dots, N$ are transmitted synchronously without distortion, propagate unaffected through space with the speed of light c , and are received without interference or distortion. Then this ideally received signal $s(t)$ is a sum of known signals with unknown phase shifts,

$$s(t) = s_1(t + \rho_1/c) + s_2(t + \rho_2/c) + \dots + s_N(t + \rho_N/c) \quad (13.11)$$

where ρ_i relate to the travel times, the geometrical ranges, and the clock offset via (Eq. 13.1) and (Eq. 13.3). Hence, to estimate the pseudoranges, the receiver has to measure the phase of the individual components. This is possible since the signals are transmitted with CDMA which makes each signal component $s_i(t)$ separable by correlating the received signal $s(t)$ with a locally generated copy of the component. In mathematical terms, the multiple access of the GNSS signals means that

$$\int_T s_i(\tau) s_j(\tau + \rho) d\tau \ll \int_T s_i(\tau) s_i(\tau) d\tau \quad \forall \rho \in \mathbb{R}, i \neq j$$

where T is a sufficiently long correlation time window. Together with the general property of the autocorrelation (Hayes 1996) of a signal,

$$\arg \max_{\rho} \int_T s_i(\tau + \alpha) s_i(\tau + \rho) d\tau = \alpha \quad (13.12)$$

where α is an arbitrary constant, this means that the pseudorange $\tilde{\rho}_i$ can be measured from the received signal $s(t)$ by

$$\tilde{\rho}_i(t) = \arg \max_{\rho} \int_T s(\tau) \bar{s}_i(\tau; \rho) d\tau \quad (13.13)$$

where the integral is the correlation of the received signal $s(t)$ with a locally generated signal component copy $\bar{s}_i(\tau; \rho) = s_i(t + \rho/c)$, parameterized with the pseudorange ρ , and T is a sufficiently long correlation window centered around t . Both in (Eq. 13.12) and in (Eq. 13.13), it has been implicitly assumed that the $\arg \max(\cdot)$ -function has been constrained to a signal period around the true delay/pseudorange or else the maximum would be ambiguous since the integrand is periodic. For further discussions about this assumption, see Sect. 4.2. In (Eq. 13.13), it has also been assumed that the pseudoranges ρ_i are essentially constant over the correlation window T .

3.2.2 Random Distortion

Now assume a random noise term $n(t)$ is present in the received signal giving a distorted signal $\tilde{s}(t)$,

$$\tilde{s}(t) = s_1(t + \rho_1/c) + s_2(t + \rho_2/c) + \dots + s_N(t + \rho_N/c) + n(t) \quad (13.14)$$

The noise can be due to nonideal transmission, refraction, interference, or measurement noise. Fortunately, as long as the noise is uncorrelated with the signal, that is,

$$\int_T n_i(\tau) s_i(\tau + \rho) d\tau \ll \int_T s_i(\tau) s_i(\tau) d\tau \quad \forall \rho \in \mathbb{R}, i = 1, \dots, N$$

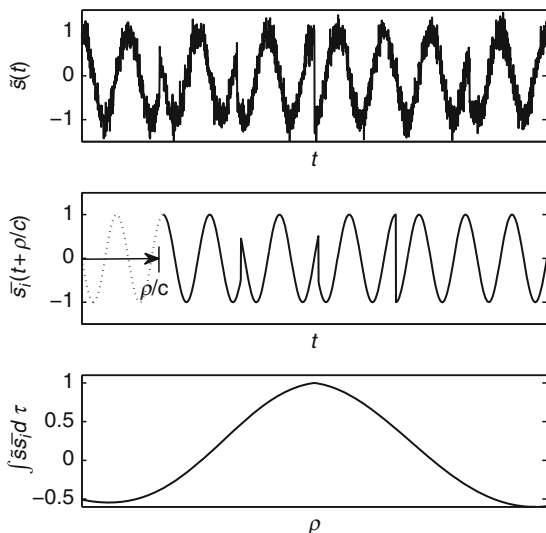
this will only cause minor errors in (Eq. 13.13), and pseudorange measurements $\tilde{\rho}_i$ can still be made by correlating the received signal with locally generated signal component copies $\tilde{s}_i(\tau; \rho)$

$$\tilde{\rho}_i(t) = \arg \max_{\rho} \int_T \tilde{s}(\tau) \tilde{s}_i(\tau; \rho) d\tau \quad (13.15)$$

A signal of the form (Eq. 13.4), illustrated in Fig. 13.6 but distorted by random noise, and the resulting correlation (Eq. 13.15) are illustrated in Fig. 13.8. Note that the correlation (Eq. 13.15) can be interpreted as a matched filter.

3.2.3 Systematic Distortion

Unfortunately, in reality, the pseudoranges are not possible to determine directly from (Eq. 13.15). The main reason for this is that there is systematic distortion in the transmission, propagation, and reception of the satellite signals, whereby a local copy of



■ Fig. 13.8

Illustration of the correlation (Eq. 13.15) (bottom) of the receiver signal $\tilde{s}(t)$ (top), consisting of a single satellite signal $s_i(t)$ and a random noise $n(t)$, with a locally generated signal component copy $\tilde{s}_i(\tau; \rho)$ (middle)

the signal components $\tilde{s}_i(\tau; \rho)$ does not necessarily have sufficient correlation with the corresponding component in the received signal. This is resolved by parameterizing the locally generated copy with respect to significant and systematic attributes of the unknown distortion and by taking the maximum correlation with respect to the phase and the parameters.

The main systematic distortion is an unknown small frequency shift due to Doppler effects caused by the satellite and the receiver moving in relation to each other. Assuming the transmitted signals to be of the form (Eq. 13.4), the received distorted signal component of satellite i can be modeled to be

$$\tilde{s}_i(t + \rho_i/c) = \sqrt{P} \cdot C_i(t + \rho_i/c) \cos\left(2\pi\tilde{f}_c(t + \rho_i/c)\right) \quad (13.16)$$

where $\tilde{f}_c = f_c + \delta f_c$, where δf_c is the unknown shift of carrier frequency. The received signal can then in turn be modeled by

$$\tilde{s}(t) = \tilde{s}_1(t + \rho_1/c) + \tilde{s}_2(t + \rho_2/c) + \cdots + \tilde{s}_N(t + \rho_N/c) + n(t) \quad (13.17)$$

Based on (Eq. 13.16), the locally generated signal components can then be parameterized as

$$\tilde{s}_i(t; \rho, \tilde{f}_c, \phi) = C_i(t + \rho/c) \cos(2\pi\tilde{f}_c t + \phi) \quad (13.18)$$

where \tilde{f}_c and ϕ is the perceived carrier frequency and phase, respectively. Note that the power P of the signal components does not affect the location of the correlation peak and has thus for simplicity been set to unity. Curiously by comparing (Eq. 13.16) and (Eq. 13.18), we see that

$$\phi = 2\pi\tilde{f}_c \cdot \rho/c \quad (13.19)$$

and consequently that the signal is over-parameterized using ρ , \tilde{f}_c , and ϕ . In this parameterization ρ corresponds to the *code phase* but the *carrier phase* ϕ could be eliminated enforcing (Eq. 13.19) and letting ρ be an overall signal phase. However, the over-parameterization will for many receivers give desirable properties. In theory, some phase information will be lost, but in practice this will help in eliminating ambiguities in the phase tracking. For further discussion on this, see Sect. 4.2.

Let the carrier parameters be $\theta = [\tilde{f}_c, \phi]$ or $\theta = \tilde{f}_c$, depending on whether (Eq. 13.19) is enforced or not, and assume that the parameters are essentially constant over the correlation window T . Then the pseudoranges can be measured from the received distorted signal $\tilde{s}(t)$ by

$$\tilde{\rho}_i(t) = \arg \max_{\rho} \left(\max_{\theta} \int_T \tilde{s}(\tau) \tilde{s}_i(\tau; \rho, \tilde{f}_c, \phi) d\tau \right) \quad (13.20)$$

In principle, this formula describes the overall function that receivers approximate when determining the pseudoranges from the raw radio measurements.

The longer the correlation window T , the more the effects of distortions with short time correlations (noise) are suppressed. Hence, T is limited by the correlation time of the parameters which in turn is often determined by the dynamics of the receiver. T can also be limited by other factors, see [◆ Sect. 4.2.4](#). In principle, the correlation could also be generalized and evaluated as a weighted integral instead of over a fixed time window.

4 Practice

In [◆ Sect. 3](#), the theory of obtaining the pseudorange measurements from the received signal and computing a receiver position estimate was discussed. In implementing this theory, a number of approximations and considerations will have to be made and taken into account. In this section, we describe implementation and related issues of the theory of [◆ Sect. 3](#). Also a brief discussion on physical receiver implementation and low-level signal processing is given.

4.1 Position Estimation

In this section, we will focus on the dominant squared error criterion for position estimation. Here, we consider the deterministic least-squares (LS) approach and the stochastic minimum mean square error (MMSE) approach for position estimation. The former assumes no prior knowledge of the motion of the receiver, while the latter assumes a dynamic model which can improve the performance especially for vehicular applications where the dynamics are greatly constrained by the vehicle.

4.1.1 LS Position Estimation

For typical error term characteristics of e_p , the minimization problem ([◆ Eq. 13.7](#)) is well behaved and a standard Gauss–Newton method can be used. The method utilizes a linearization of the measurement residuals ([◆ Eq. 13.8](#)) around parameter estimates $[\hat{x}_j, \hat{y}_j, \hat{z}_j, \delta \hat{t}_j] = \hat{\mathbf{x}}_j$ at iteration j ,

$$r_i(\mathbf{x}) \simeq r_i(\hat{\mathbf{x}}_j) + \frac{\partial r_i}{\partial [x, y, x, \delta t]} \bigg|_{\hat{\mathbf{x}}_j} \Delta \mathbf{x}_j \quad (13.21)$$

where \simeq denotes equality, ignoring higher order terms, and where $\Delta \mathbf{x}_j = [\Delta x_j, \Delta y_j, \Delta z_j, \Delta \delta t_j]^T$ are local coordinates around $\hat{\mathbf{x}}_j$,

$$\Delta \mathbf{x}_j = \mathbf{x} - \hat{\mathbf{x}}_j \quad (13.22)$$

where \mathbf{x} are the absolute coordinates in use. Stacking all linearized residual relations ([◆ Eq. 13.21](#)),

$$\underbrace{\begin{bmatrix} r_1(\mathbf{x}) \\ r_2(\mathbf{x}) \\ \vdots \\ r_N(\mathbf{x}) \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} r_1(\hat{\mathbf{x}}_j) \\ r_2(\hat{\mathbf{x}}_j) \\ \vdots \\ r_N(\hat{\mathbf{x}}_j) \end{bmatrix}}_{\mathbf{b}_j} + \underbrace{\begin{bmatrix} \frac{\partial r_1}{\partial [x, y, z, \delta t]} \Big|_{\hat{\mathbf{x}}_j} \\ \frac{\partial r_2}{\partial [x, y, z, \delta t]} \Big|_{\hat{\mathbf{x}}_j} \\ \vdots \\ \frac{\partial r_N}{\partial [x, y, z, \delta t]} \Big|_{\hat{\mathbf{x}}_j} \end{bmatrix}}_{\mathbf{A}_j} \underbrace{\begin{bmatrix} \Delta x_j \\ \Delta y_j \\ \Delta z_j \\ \Delta \delta t_j \end{bmatrix}}_{\Delta \mathbf{x}_k} \quad (13.23)$$

we have an equation system of the form,

$$\mathbf{r} = \mathbf{b}_j + \mathbf{A}_j \Delta \mathbf{x}_j \quad (13.24)$$

Then, given the linearization point $\hat{\mathbf{x}}_j$, $\|\mathbf{r}\|^2$, equivalent to the minimization criterion of (Eq. 13.9), is minimized by

$$\Delta \mathbf{x}_j^{\min} = -(\mathbf{A}_j^T \mathbf{A}_j)^{-1} \mathbf{A}_j^T \mathbf{b}_j. \quad (13.25)$$

updating the position estimate (linearization point) by

$$\hat{\mathbf{x}}_{j+1} = \hat{\mathbf{x}}_j + \Delta \mathbf{x}_j^{\min} \quad (13.26)$$

and iterating (Eq. 13.25) and (Eq. 13.26), an approximation to (Eq. 13.9) is attained. The initial estimate $\hat{\mathbf{x}}_0$ is often taken to be the center of the Earth. The process normally converges to meter-level accuracy within two to three iterations (Borre et al. 2007). Let m denote the iteration at which the process is terminated.

Naturally the accuracy of the final position estimate $\hat{\mathbf{x}}_m$ is dependent on the accuracy of the pseudorange measurements but also on the geometrical distribution of the satellites. Assume that a confidence interval of the actual pseudorange is given around the measured pseudorange. Then as illustrated in two dimensions in Fig. 13.4, the volume containing the true position will not be equally distributed around the mean, and the position error might even be larger than the individual errors in the pseudoranges. This effect on the precision due to the geometry of the satellites is called dilution of precision (DOP). In loose terms, the position DOP is the factor by which the pseudorange errors are amplified when position is estimated based on them. As a by-product of the iteration (Eq. 13.25) and (Eq. 13.26), an estimate of the DOP is given by the error covariance of the minimizing argument of (Eq. 13.25). Assuming zero-mean errors with equal covariances σ_ρ^2 for all the pseudorange measurements, the error covariance of the estimate $\text{cov}(\mathbf{x}^* - \hat{\mathbf{x}}_m)$ can be approximated by

$$\text{cov}(\mathbf{x}^* - \hat{\mathbf{x}}_m) = \sigma_\rho^2 \frac{1}{N-4} (\mathbf{A}_m^T \mathbf{A}_m)^{-1} \quad (13.27)$$

where x^* denotes the true position value. Typical values for the position DOP are below 2 (95% of the time). In combination with errors as discussed in [Sect. 4.3](#), typical position errors are below 10 [m] (95% of the time) (Kaplan and Hegarty 2006).

4.1.2 MMSE Position Estimation

For vehicular applications, the dynamics of the receiver are constrained, with significant correlations for position, velocity, and acceleration in time. This prior information can be exploited by tracking the moving receiver instead of solving for the position at every time instant as of [Sect. 4.1.1](#). The dynamics can be modeled by,

$$\mathbf{z}_{k+1} = \mathbf{F}\mathbf{z}_k + \mathbf{G}\mathbf{w}_k \quad (13.28)$$

for discrete time instants k , where the vector \mathbf{z} contains the receiver and vehicle states, for example, the position, velocity, and the receiver clock offset, and the system matrix \mathbf{F} describes their functional dependence. Changes in the states are modeled as random, memoryless processes, captured in the white process noise \mathbf{w}_k with a given covariance matrix $\mathbf{Q}_k = \text{cov}(\mathbf{w}_k)$ and a transfer matrix \mathbf{G} . Nonlinear dynamic models are also possible to use instead of ([Eq. 13.28](#)) at higher computational cost.

The pseudorange measurements $\tilde{\rho}_{i,k}$ from the visible satellites i at time instant k are treated as observations modeled as functions of the states \mathbf{z} ,

$$\tilde{\rho}_{i,k} = \sqrt{(x_k - x_{i,k})^2 + (y_k - y_{i,k})^2 + (z_k - z_{i,k})^2} + c\delta t_k + e_{i,k}^{\text{model}} + \eta_k \quad (13.29)$$

where η_k is random measurement noise with a covariance matrix \mathbf{R}_k . Using the dynamic model ([Eq. 13.28](#)) and the measurement model ([Eq. 13.29](#)), the tracking problem can be solved recursively by, for example, an extended Kalman filter (EKF), which is designed to minimize the MSE of an estimate of \mathbf{z}_k in the linearized system. This in turn giving an estimate of ([Eq. 13.10](#)) (Kailath et al. 2000; Farrell 2008).

Analogous to the DOP estimate in the LS-approach, the EKF provides an approximate error covariance matrix $\text{cov}(\mathbf{z}_k^* - \hat{\mathbf{z}}_k)$, where \mathbf{z}_k^* denotes the true and $\hat{\mathbf{z}}_k$ denotes the estimated system state, for each time instant k , from which a measure of the estimator's uncertainty of the position can be extracted.

4.2 Measuring Pseudoranges

From an implementation point of view, there are a number of difficulties with ([Eq. 13.20](#)). The two most important ones are the extent of the search space and spatial ambiguities.

The extent of the search space when searching for the maximum of ([Eq. 13.20](#)) is too large for it to be feasible to make a global search for every time instant. Also, due to

noise there will always be maxima in the integral of (Eq. 13.20) even if a satellite signal is not received due to, for example, blockage. Such maxima do not correspond to actual pseudoranges and should therefore be ignored. Fortunately, the pseudoranges, which essentially determine the maxima, are strongly correlated in time. Hence, a maximum can be tracked rather than solved for at each time instant. For this purpose, phase-lock loops (PLL) and delay-lock loops (DLL) are typically employed. However, the loops need to be initialized with values corresponding to a received signal within the pull-in range of the loops. The search for such values is called *signal acquisition*.

The spatial ambiguities have to do with nonunique correlation maxima. As noted in Sect. 4.2, the integral of (Eq. 13.20) is periodic and will have maxima spaced with the period of the signal. For the total period of the signal, that is, the period of the code $C_i(t)$ typically being > 1 [ms], this is not a problem since all but one correlation maxima will indicate positions far from the surface of the Earth. However, the period of the carrier is much shorter (equivalent to submeter) and will cause closely spaced local maxima to dominate the integral. Typically, the available information is not sufficient to resolve the local maxima creating ambiguities and resulting in erratic behavior of a naïve implementation of the lock loops. Consequently, the lock loops typically have to be constructed to be insensitive to the absolute carrier phase ϕ and to track the carrier separately from the code phase, hence the redundant separate carrier and code parameterization in (Eq. 13.20).

The maximization with respect to the carrier parameters ϕ and \tilde{f}_c will pick a maximum independent of the code phase and as long as this maximum is accurately pinpointed, the maximizing code phase argument ρ will be independent of the carrier phase ϕ . On the other hand, if sufficient information is available to resolve the local maxima, improved accuracy can be expected. If carrier phase information is used, one speaks about *carrier phase tracking*.

4.2.1 Signal Acquisition

Signal acquisition is the process of determining the presence of a satellite signal and finding values for initializing the lock loops as well as reinitializing them if the locks have been lost.

Due to the spatial ambiguities and to limit the dimensionality of the search space, the received signal is typically transformed to eliminate the sensitivity to carrier phase. The signal acquisition is then the search for which satellites i are present, along with their individual pseudoranges ρ and carrier frequencies \tilde{f}_c , using a carrier phase insensitive equivalent of the integral in (Eq. 13.20). The simplest but computationally least efficient way to achieve such an equivalent of (Eq. 13.20) is to use the magnitude of the in-phase (I) and quadrature (Q) components of the received signal. This IQ-decomposition gives the search function as

$$\arg \max_{\rho, \tilde{f}_c} \left(\left| \int_T \overbrace{\tilde{s}(\tau) \tilde{s}_i(\tau; \rho, \tilde{f}_c, \phi)}^{\text{I-component}} d\tau \right|^2 + \left| \int_T \overbrace{\tilde{s}(\tau) \tilde{s}_i(\tau; \rho, \tilde{f}_c, \phi + \pi/2)}^{\text{Q-component}} d\tau \right|^2 \right) \quad (13.30)$$

where $\tilde{s}_i(\tau; \rho, \tilde{f}_c, \phi)$ and $\tilde{s}_i(\tau; \rho, \tilde{f}_c, \phi + \pi/2)$ correspond to two locally generated signal components with the phase of the carrier shifted by 90° . Ideally this expression becomes independent of ϕ , and the maxima with respect to ρ and \tilde{f}_c are identical to those of (Eq. 13.20).

The search for the maxima is straightforward. The search space of the parameters is discretized with sufficient resolution, and an exhaustive search is conducted. Other carrier phase insensitive equivalents of (Eq. 13.30) giving numerically more efficient but also more complex search strategies exist. These search strategies perform parallel searches in the parameter space by exploiting properties of the Fourier transform.

4.2.2 Carrier and Code Phase Tracking

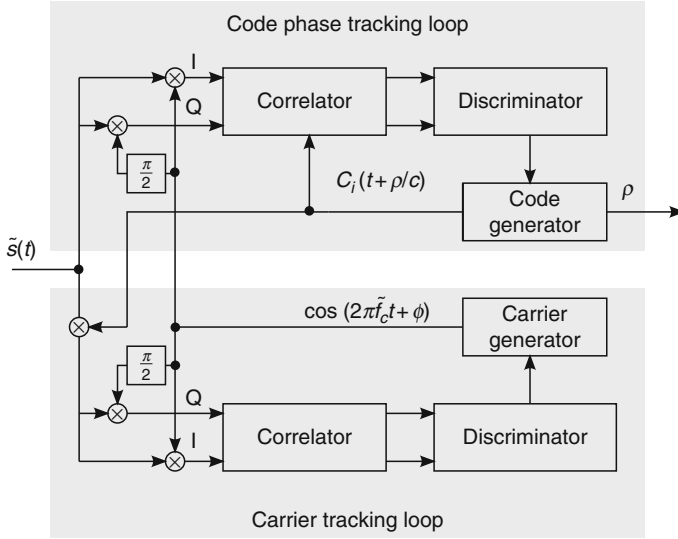
Given initial values of ρ and \tilde{f}_c from the signal acquisition process, the receiver needs to find and track the maximum of (Eq. 13.20). This is performed with two interleaved tracking loops. As pointed out before, the carrier phase cannot typically be resolved, and therefore the carrier (\tilde{f}_c and ϕ) is tracked separately from ρ . To make the tracking loops insensitive to carrier phase errors, similar transformation as of (Eq. 13.30) is used for the tracking loops. In general, the loops use separate correlators for the I- and Q-components and apply feedback based on error discriminators applied to the correlator outputs.

The two lock loops of the receiver perform in parallel, with one loop using the tracked parameter value of the other loop,

$$\left[\phi_i(t), \tilde{f}_{c,i} \right] = \arg \max_{\phi, f} \int_{T_\phi} \tilde{s}(\tau) \tilde{s}_i(\tau; \rho_i, f, \phi) d\tau \quad (13.31)$$

$$\rho_i(t) = \arg \max_{\sigma} \int_{T_\rho} \tilde{s}(\tau) \tilde{s}_i(\tau; \sigma, \phi_i(t), \tilde{f}_{c,i}) d\tau \quad (13.32)$$

where T_ϕ and T_ρ are time windows of the correlators centered around t . This functionality is typically implemented as illustrated in Fig. 13.9. The carrier tracking (Eq. 13.31) is typically performed with a PLL and the code tracking (Eq. 13.32) with a DLL. Note that the phase $\phi_i(t)$ of the carrier tracking loop corresponding to (Eq. 13.31) is ambiguous due to the short period of the carrier. However, any maximum will do since it wipes off the carrier equally well in (Eq. 13.32) making the code tracking loop insensitive to absolute carrier phase. Consequently, the tracked phase $\phi_i(t)$ cannot be used for ranging and instead the output of the tracking loop corresponding to (Eq. 13.32) gives the measured pseudorange.



■ Fig. 13.9

Implementation of the pseudorange ρ_i extraction based on code phase only. The carrier tracking loop tracks the carrier component of the signal, and the code phase tracking loop tracks the phase of the code, which gives a measurement of the pseudorange. One such pair of tracking loops is used for each tracked satellite signal

An intuitive description of loops can be attained by noting that (● Eq. 13.31) and (● Eq. 13.32) essentially become

$$\left[\phi_i(t), \tilde{f}_{c,i} \right] = \arg \max_{\varphi, f} \int_{T_\phi} \cos(2\pi \tilde{f}_c \tau + \phi) \cos(2\pi f \tau + \varphi) d\tau \quad (13.33)$$

$$\rho_i(t) = \arg \max_{\sigma} \int_{T_p} C_i(t + \rho/c) C_i(t + \sigma/c) d\tau \quad (13.34)$$

Hence, ideally the loops maximize the (cross-)correlation of the corresponding received and locally generated information signal and carrier.

4.2.3 Carrier Phase Tracking

As noted, the over-parametrization of the received signal is beneficial in the sense that it removes correlation maxima ambiguities. On the other hand, this also means that the phase information in the carrier is lost. If the carrier phase is tracked and information is available such that the maxima can be resolved, this can be expected to result in improved positioning performance. Resolving the maxima consists of determining which of the correlation maxima that corresponds to the true carrier phase in relation

to the code phase. Note that the code and the carrier are synchronized upon transmission. A code period consists of an even number of carrier periods. Resolving the maxima typically requires good signal reception and some system augmentation, see ➊ Sect. 5. Once the maxima have been resolved, the resulting carrier phase can be tracked by an ordinary carrier tracking loop initialized with the result of the resolution. Yet providing better position accuracy, carrier phase receivers are also typically much more expensive than code phase the like.

4.2.4 Navigation Data

In the outlined pseudorange measurement methods, it has been assumed that the transmitted information signals $C_i(t)$ are perfectly known. For some GNSS signals, this is the case but for the commonly used ones it is not. As noted in ➊ Sect. 2.2, the positions of the satellites are commonly encoded in the signals. This means that the information signal $C_i(t)$ contains a uniquely identifiable periodic signature or code $C'_i(t)$ but also an unknown overlaid bipolar navigation message $D_i(t)$ containing information about the orbit (position) of the satellite,

$$s_i(t) = C_i(t) \cos(2\pi f_c t) = \underbrace{D_i(t)}_{\text{unknown}} \cdot \underbrace{C'_i(t) \cos(2\pi f_c t)}_{\text{known}} \quad (13.35)$$

This unknown component $D_i(t)$ will cause some limitations, but the pseudorange measurements are still possible. The bipolar nature of $D_i(t)$ causes an unknown $\pm 180^\circ$ phase shift of the known signal but the unknown $D_i(t)$ -modulation is significantly slower than, and synchronized with, the period of the known $C'_i(t)$. This means that by constructing discriminators that are insensitive to these $\pm 180^\circ$ phase shifts the lock loops will perform as expected in between the phase shifts. Then by synchronizing the correlation windows with the phase shifts based on the estimated signal phase (pseudoranges) and using integrate-and-dump in the correlators (in contrast to sliding windows), the effect of the unknown navigation message $D_i(t)$ is suppressed. Finally, by wiping off the known signature $C'_i(t)$ and the carrier $\cos(2\pi f_c t)$ from the received signal, the navigation message can be extracted and the position of the satellite calculated.

4.3 Pseudorange Error Sources

The error term e_i in the pseudorange relation (➊ Eq. 13.5) accounts for the nonideal signal transmission, propagation, and reception of the signal from satellite i and imperfections in the tracking loops resulting in an error in the pseudorange measurement. The term e_i can be broken down in components with different levels of details. A rough division is: satellite (nonideal transmission)-related errors e_i^{sat} ; atmospheric propagation-related errors typically divided into ionospheric and tropospheric propagation errors, d_i^{ion}

and d_i^{trop} ; local propagation path (multipath)-related errors e_i^{mp} ; and interference, receiver-related (nonideal reception), and residual errors n_i ,

$$e_i = e_i^{\text{sat}} + d_i^{\text{ion}} + d_i^{\text{trop}} + e_i^{\text{mp}} + n_i \quad (13.36)$$

The nonideal signal transmission e_i^{sat} arises mainly from satellite clock errors, imperfect ephemeris (orbital) information, and relativistic effects, resulting in a phase shift and an equivalent range error. This term can be mitigated by error models and correction data provided by an augmentation system.

Nonideal signal propagation occurs as the electromagnetic waves from a satellite penetrates the atmosphere, which consists of layers of gases that modify the signal properties, in particular lower electromagnetic propagation speed through the medium. This lower propagation speed gives rise to delays at the receiver that have to be compensated for by atmospheric models and external monitoring information provided to the receiver. The gases of the outermost part of the atmosphere, at altitudes above 80 [km], called the *ionosphere*, are ionized by solar radiation into a plasma of free electrons and ions. This causes a frequency-dependent refractive index and a lower propagation speed as in comparison to that of vacuum, c , which introduces a delay and an equivalent range error d_i^{ion} which varies with frequency and the model must compensate for variations in space and time, including effects from sunspot cycles. Using multiple signals with different frequencies, this delay can to some extent be measured and compensated for. Also models and data from augmentation systems can be used to compensate for the delay. The *troposphere* is the innermost part of the atmosphere and extends up to about 10 [km]. The refractive index of the troposphere depends on temperature, pressure, and humidity. The resulting delay and equivalent range error d_i^{trop} can to some extent be compensated for by models and data from an augmentation system.

Local path errors can occur through reflection before the signal reaches the receiver, for example, in urban areas. Multiple reflections will by themselves extend the propagation path beyond the direct path, but can further interfere with a directly received signal component. This results in so-called multipath effects whereby the receiver finds erroneous correlation peaks and hence biased pseudorange estimates, all of which are summarized in e_i^{mp} . In general, compensating these effects is a difficult signal processing problem and is a common cause of failure of satellite navigation systems in dense urban environments. To some extent, the multipath can be mitigated by directional antennas or by aiding data for the correlators.

Nonideal signal reception is mainly due to the interference, receiver noise, receiver quantization, and component nonlinearities. These effects in combination with residual errors will typically be treated as random noise n_i on the pseudorange measurements.

Rough figures of the magnitude of the different error sources are given in [Table 13.2](#), for standard single-frequency code phase receivers after compensations by standard error models but without the use of augmentation data, cf. (Tsui 2005; Kaplan and Hegarty 2006; Grewal et al. 2007). In open-sky unobstructed environments, the satellite and atmospheric errors will normally dominate, while in obstructed environments such as indoor and urban canyons, multipath errors can dominate the error in position. Note that

■ **Table 13.2**
Rough error budget of pseudorange measurements

Error source	Term	1- σ error [m]
Satellite	e_i^{sat}	1.5–4
Ionosphere	d_i^{ion}	5–10
Troposphere	d_i^{trop}	0.2–3
Multipath	d_i^{mp}	0.2–25
Receiver and interference	n_i	0.1–2

in unobstructed environments, the equivalent error in position is typically somewhat smaller than expected from the atmospheric error figures and normal DOP values. This is due to the fact that atmospheric errors are correlated between different satellites and as noted, common mode errors do not affect the position solution (► Eq. 13.9).

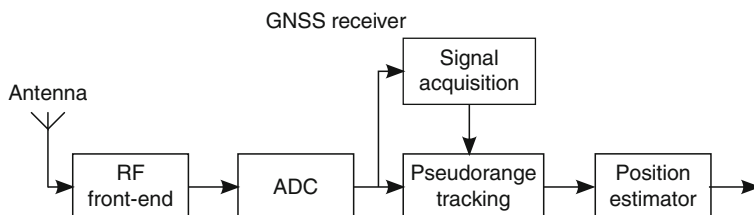
4.4 GNSS Receivers

The position estimation algorithm of choice, the signal acquisition, and the pseudorange measurement lock loops need to be physically implemented in a GNSS receiver. Apart from this higher-level signal processing, the receiver implementation also has to include the lower-level processing to produce a suitable signal for the higher levels of processing. The transmitted signal needs to be received by an antenna and processed by a radio frequency (RF) front end and typically also sampled by an analog-to-digital converter (ADC). A generic division of the receiver into functional blocks is illustrated in ► Fig. 13.10. In this section, some further comments about the different blocks and some other receiver-related issues, will be given.

The range of GNSS receivers is large. In the low end, there are single-chip receivers where all the functions of ► Fig. 13.10 have been implemented on a single application-specific integrated circuit (ASIC), while in the high end, there are receivers in which each component is individually manufactured, calibrated, and assembled. Another extreme is the software receivers where a majority of the functionalities are implemented in software. The range in price is also large, from tens of Euros for the low-end receivers to tens of thousands of Euros for high-end receivers. A good overview of the the GNSS receiver market is found in the annual GPS World Receiver Survey (World Staff 2010b).

4.4.1 GNSS Antennas

A wide variety of GNSS antennas are available on the market. In the low end, there are typically antennas printed on circuit boards and possibly integrated into the receiver assembly. For mid-range and high-end receivers, there are a range of antennas, with



■ Fig. 13.10

Generic division of GNSS receiver functionality

increasingly complex geometrical shape and cost but also with increasing efficiency, as well as active antennas and antenna arrays for interference and multipath mitigation. The important parameters for GNSS antennas are *center frequency*, *bandwidth*, *polarization*, and *gain pattern*. For external antennas, a preamplifier is often integrated with the antenna to mitigate noise sensitivity and allow for long antenna cables. A practical overview over the antenna market can be found in the annual GPS World Antenna Survey (World Staff 2010a).

4.4.2 RF Front End and ADC


The GNSS RF front end and ADC perform the traditional tasks of bandpass filtering, amplification, mixing the signal down to intermediate frequency, and finally sampling the signal. However, the situation for the GNSS front end and ADC is in some aspects quite different from that of a more traditional non-spread-spectrum communication RF front ends and ADCs. First, the power of the received signal is below the thermal noise floor in a frequency bands $\sim 2\text{--}40$ [MHz] wide in the $1.2\text{--}1.6$ [GHz] region. Therefore, the requirements on the amplification and the bandpass selectivity are quite extreme. The amplification should be in the order of 100 [dB] over a frequency band with a width of about 0.1% of the center frequency. To handle this, the GNSS RF front ends are normally designed to perform the filtering, amplification, and mixing in several stages. Second, the main information of interest in the signal is not the individual code bits but correlation peaks with long known code sequences. Therefore, the analogue-to-digital conversion can be performed with very few bits. In low-cost receivers, a single-bit ADC is often used. An advantage of this is that no automatic gain control is needed. In high-end receivers, up to 3-bits ADCs are used.

4.4.3 Hardware Signal Acquisition Implementation

Typically, the signal acquisition block, performing the search of (Eq. 13.30), is implemented together with the pseudorange tracking block in a single ASIC. To speed up the signal acquisition and reacquisition, current receivers are normally equipped with multiple parallel search channels. Thus, a single channel can continuously search for the

presence of a satellite signal. This simplifies receiver design because each pseudorange tracking loop can be hardwired to an acquisition channel. If this is not the case, the acquisition channels will have to be time multiplexed with respect to the different presumed satellite signals.

4.4.4 Hardware Pseudorange Tracking Loop Implementation

Typically, the pseudorange tracking block is implemented together with the signal acquisition block in a single ASIC. One tracking channel, comprising an implementation of the tracking loops illustrated in  Fig. 13.9 or equivalent, has to be present for each satellite to be tracked. Hence, a minimum of four tracking channels are needed but current receivers typically use eight or more channels, sufficient to track all satellites in view.

4.4.5 Position Estimation

The position estimation algorithms are typically implemented, together with error models, ephemeris models, and augmentation system correction capabilities, on a microcontroller integrated in the receiver. For system integration, the position estimation can also be left to other agents by letting the receiver output only pseudorange measurements.

4.4.6 Software Receivers

During the last decade, there has been an increasing interest in trying to transfer as large portion as possible of the signal processing from hardware to software. Such receivers are typically distinguished as software receivers. Primarily this concerns the signal acquisition and pseudorange tracking which in contrast to the position estimation is normally performed in dedicated hardware. The advantage with transferring functionality from hardware to software is the reduced amount of needed hardware and the gain in flexibility, both as of signal processing methods and as of integration with other systems. The difficulty with the transfer is the large computational burden. Also, currently as stand-alone GPS receivers, the hardware receivers are typically less expensive and consume less power than a software the like. However, with an expanding toolbox of signal processing methods and the availability of complementary sensors in a vehicle, software receivers are an attractive alternative for many vehicular applications.

5 Differential GNSS

For many outdoor GNSS user scenarios, the dominating error sources of the tracked pseudoranges are atmospheric (d_i^{ion} and d_i^{trop}) and satellite (e_i^{sat}) related errors. These errors are strongly correlated over large areas to a decreasing degree from kilometers to the

entire globe. Hence, they can be measured by external so-called augmentation systems and compensated for in the receiver. The technique of measuring and compensating for such errors are referred to as differential GNSS (DGNSS). Also, internal augmentation based on additional sensors, maps, etc., in a wider sensor fusion system is possible (Farrell 2008). However, here we will focus on the external augmentation systems.

A DGNSS augmentation system consists of reference stations with well-surveyed positions, together with a processing center and a system to convey the correction information. Based on its known position, the monitoring stations can measure errors in the received satellite signals and monitor the health and integrity of the satellites. The measured errors are processed and used to correct for the errors experienced by closely located GNSS receivers.

With correction from an augmentation system, a GNSS receiver can typically achieve horizontal position accuracies of >1 [m] for code phase receivers and >0.1 [m] for carrier phase receivers. This makes DGNSS systems highly interesting for vehicle applications since it allows the vehicle to be positioned in a specific lane.

5.1 Principles

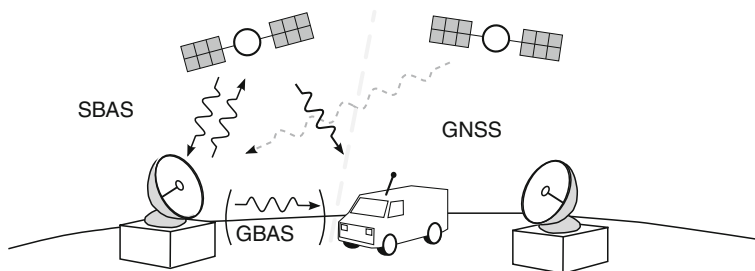
In a basic setup, the reference stations measure the pseudoranges $\tilde{\rho}_i^{\text{ref}}$ and differentiate them with ranges $\gamma_i^{\text{ref}}(x, y, z)$ calculated from its known position. Due to the ground station location and receiver quality, path errors e_i^{path} and nonideal reception errors n_i can typically be neglected. Consequently, the differences give a measure of the error terms (🔗 Eq. 13.36) apart from the augmentation system clock offset δt^{ref} ,

$$\tilde{\rho}_i^{\text{ref}} - \gamma_i^{\text{ref}}(x, y, z) \approx c\delta t^{\text{ref}} + e_i^{\text{sat}} + d_i^{\text{ion}} + d_i^{\text{trop}} \quad (13.37)$$

Using measurements from multiple satellites, measurements of signals at different frequencies, measurements from several reference stations, and error models, the receiver clock error can be estimated and the error components separated. For corrections in a local area, the difference (🔗 Eq. 13.37) can be used to correct the pseudoranges measured by nearby receivers. For corrections over larger areas, the error components need to be separated, and models to interpolate and extrapolate the components need to be used. Such processing of the error measurements is typically done centrally in the augmentation systems. The corrections provided by the augmentation systems can either be sent directly as pseudorange corrections or as parameters in error models.

5.2 Augmentation Systems

Based on the distribution system, the augmentation systems can be divided into satellite-based augmentation systems (SBAB) and ground-based augmentation systems (GBAS). The SBABs broadcast the information via a network of geostationary satellites, while the GBASs broadcast the information from ground-based transmitters. The SBABs typically



■ Fig. 13.11

Augmentation system infrastructure in addition to the existing GNSS infrastructure as of [Fig. 13.7](#). The augmentation system requires additional ground stations, and in the case of SBAS additional satellites. The SBAS broadcasts the augmentation information to the user via satellites (arrow from satellite to vehicle) while the GBAS broadcasts the augmentation information from ground stations (arrow from ground station to vehicle). In the latter case, no space segment is present

have large coverage areas, from countries to continents to the entire globe. The GBASs typically cover smaller areas from individual sites to regions and countries. The additional infrastructure to the existing GNSS infrastructure and the information transfer within the systems are illustrated in [Fig. 13.11](#).

5.2.1 Satellite-Based Augmentation Systems

The SBASs often broadcast their correction information encoded in signals with identical structure as of the GNSS signals. Hence, typically only minor modifications of GNSS receivers are needed to receive the correction information.

There are currently a number of SBASs available or under construction. These systems can be divided into regional government-supported and private (entirely commercial) systems. The government-supported systems cover corresponding regions and are designed and controlled for interoperability and to provide a seamless coverage. They also provide open access to at least a subset of the broadcasted signals and supply corrections that can be used for a wide range of standardized receivers. The commercial systems on the other hand have essentially global coverage but require subscriptions and special purpose receivers.

There are three operational and two additional planned government-supported systems. Covering North America, apart from southern Mexico and parts of the Arctic regions, is the GPS *Wide Area Augmentation System* (WAAS). Covering primarily Europe but also Northern Africa and parts of the Middle East and India is the *European Geostationary Navigation Overlay Service* (EGNOS). Work to extend the coverage to Southern Africa has commenced. The system is intended to augment the GPS, the GLONASS, and the Galileo systems but is currently only supporting augmentation of

■ Table 13.3
Summary and currently operational and planned SBASs

SBAS	Operator	Coverage	Accuracy [m]	Open access
WAAS	USA, Canada	North America	<1.5	Yes
EGNOS	EU	Europe (Africa, Middle East)	<1.5	Yes
MSAS	Japan	Japan	<2	Yes
GAGAN*	India	India	–	Yes
SNAS*	China	China	–	Yes
QZSS*	Japan	Japan	–	Yes
StarFire	John Deer	Global	~0.1	No
OmniSTAR	OmniSTAR (Fugro)	Global	0.1–1	No

*System not yet operational

the GPS system. Covering Japan is the Japanese *Multifunctional Satellite Augmentation System* (MSAS). Under development for coverage of India and China, respectively, are the Indian *GPS Aided Geo Augmented Navigation* (GAGAN) system and the Chinese Satellite Navigation Augmentation System (SNAS). Japan is also currently constructing the Quasi-Zenith Satellite System (QZSS) which among other capabilities have the role of an SBAS. Two private SBAS are currently operational, the StarFire and OmniSTAR systems. A summary of currently operational and planned SBASs is found in the ● Table 13.3. The accuracy given in the table is only rough figures and intended to give an idea of the system capabilities. The lower accuracy figure for the private systems is due to special purpose carrier phase receivers.

5.2.2 Ground-Based Augmentation Systems

GBAS systems are typically deployed in local areas in which there is a need for precise position information. For example, the Local Area Augmentation System (LAAS) is installed at a number of airports to supply augmentation information as an instrumentation landing system. Similarly, the US Maritime Differential GPS (MDGPS) system and the European Differential Beacon Transmitters are a network of base stations and augmentation system transmitters situated along the costal waters of the USA and Europe. The former has also been expanded into the US Nationwide Differential GPS (NDGPS).

6 Conclusion and Further Reading

In this chapter, an overview of the principles, the theory, and practical aspects of the GNSS technology has been presented. GNSSs can provide a global position capability for vehicles and is identified as an enabling technology for many intelligent vehicle

capabilities. The positioning of the GNSS is made possible by geometrical relations between satellites at known locations and measurements of the phase (pseudorange) of the synchronously transmitted satellite signals. The phase of the signal is measured by correlating the received signal with locally generated copies of the individual satellite signals. Based on the pseudorange measurements, the position is estimated by multilateration in four or more dimensions. A host of error sources affect the pseudorange measurements and need to be compensated for by error models. In unobstructed environment, the residual errors are dominated by satellite- and atmosphere-related errors, without additional information the position accuracy is typically < 10 [m] 95% of the time. Errors in the pseudoranges can be mitigated by augmentation systems achieving sub-meter accuracy. Multiple augmentations systems are available.

The treatment of the GNSS technology in this chapter is by no means exhaustive. Several running meters of literature exist on the subject and related areas. Two good general references are (Hofmann-Wellenhof et al. 2007) and (Kaplan and Hegarty 2006). Most books on the subject will cover the basics but have a specific focus area. Discussion and information about antenna design in general is found in (Balanis 2005) and specifically about GNSS antenna design for vehicles can be found in (Rabinovich et al. 2010). Hardware receiver implementation is discussed in (Tsui 2005; Borre et al. 2007; Samper et al. 2009). Software receiver implementations are discussed in (Tsui 2005; Borre et al. 2007; Pany 2010). Finally, discussion on integration of GNSS with other types of navigation system and navigation data can be found in several books, for example, in (Grewal et al. 2007; Farrell 2008; Groves 2008; Gleason and Gebre-Egziabher 2009).

For in deeper studies of GNSS, a solid understanding of the fundamental statistical signal processing, filtering, and estimation theory is needed. Good starting points for such studies are (Hayes 1996; Oppenheim et al. 1997; Kay 1998; Kailath et al. 2000).

References

-
- | | |
|--|--|
| Balanis C (2005) Antenna theory: analysis and design. Wiley, Hoboken | GNSS technology and applications series. Artech House, Boston |
| Borre K, Akos DM, Bertelsen N, Rinder P, Jensen SH (2007) A software-defined GPS and Galileo receiver: a single-frequency approach. Birkhäuser, Boston | Hayes M (1996) Statistical digital signal processing and modeling. Wiley, New York |
| Farrell JA (2008) Aided navigation: GPS with high rate sensors. McGraw-Hill, New York | Hofmann-Wellenhof B, Lichtenegger H, Wasle E (2007) GNSS-global navigation satellite systems: GPS, GLONASS, Galileo, and more. Springer, Vienna |
| Gleason S, Gebre-Egziabher D (2009) GNSS applications and methods, GNSS technology and applications. Artech House, Boston | Kailath T, Sayed AH, Hassibi B (2000) Linear estimation. Prentice Hall, Upper Saddle River |
| Grewal MS, Weill LR, Andrews AP (2007) Global positioning systems, inertial navigation, and integration. Wiley, Hoboken | Kaplan E, Hegarty C (2006) Understanding GPS: principles and applications, Artech House mobile communications series. Artech House, Boston |
| Groves P (2008) Principles of GNSS, inertial, and multi-sensor integrated navigation systems, | Kay S (1998) Fundamentals of statistical signal processing: estimation theory, Prentice Hall signal processing series. Prentice-Hall PTR, Englewood Cliffs |

- Oppenheim A, Willsky A, Nawab S (1997) Signals and systems, Prentice Hall signal processing series. Prentice Hall, Upper Saddle River
- Pany T (2010) Navigation signal processing for GNSS software receivers, GNSS technology and applications series. Artech House, Norwood
- Proakis J, Salehi M (2008) Digital communications, McGraw-Hill higher education. McGraw-Hill, Boston
- Rabinovich V, Alexandrov N, Alkhateeb B (2010) Automotive antenna design and application. Taylor and Francis, Boca Raton
- Samper J, Pérez R, Lagunilla J (2009) GPS & Galileo: dual RF front-end receiver and design, fabrication, and test, Communication engineering. McGraw-Hill, New York
- Grelier T, Dantepal J, Delatour A, Ghion A, Ries L (2007) Initial observation and analysis of Compass MEO satellite signals. Inside GNSS:39–43
- Tsui J (2005) Fundamentals of global positioning system receivers: a software approach, Wiley series in microwave and optical engineering. Wiley, New Jersey
- GPS World Staff (2010) GPS world antenna survey. GPS World (Feb):37–49
- GPS World Staff (2010) GPS world receiver survey. GPS World (Jan):35–56
- Gao GX, Chen A, Lo S, de Lorenzo D, Walter T, Enge P (2007) Compass-M1 broadcast codes and their application to acquisition and tracking. Inside GNSS:36–43

14 Enhancing Vehicle Positioning Data Through Map-Matching

Mohammed A. Quddus¹ · Nagendra R. Velaga²

¹Department of Civil and Building Engineering, Loughborough University, Leicestershire, UK

²ITS, dot.rural Digital Economy Research Hub, University of Aberdeen, UK

1	<i>Spatial Road Map</i>	344
1.1	Errors in Map Creation	344
1.2	Errors in Digitization	345
1.3	Numerical Data Related to a Spatial Road Map	346
2	<i>Digital Map Quality</i>	346
3	<i>Methods of Map-Matching</i>	349
3.1	Geometric Analysis	350
3.2	Topological Analysis	350
3.3	Analysis with Advanced Techniques	353
4	<i>Integrity and Reliability of Map-Matching</i>	356
5	<i>Intelligent Vehicles and Map-Marching Algorithms</i>	360
6	<i>Summary</i>	361

Abstract: In-vehicle navigation systems usually rely on the integration of data from a range of positioning sensors/systems such as GPS or GPS integrated with other positioning sensors. Even with very robust sensor calibration and sensor fusion methods, positioning inaccuracies are sometimes unavoidable. In addition, there are inaccuracies with a digital road map due to errors in map creation, projection and digitization. As a result of such imprecision in the positioning systems and the faulty digital base map, actual vehicle positions do not always match with the spatial road map although the vehicle is known to be restricted on the road network. This phenomenon is referred to as spatial mismatch. The spatial mismatch is often more severe at junctions, roundabouts, complicated flyovers and built-up urban areas with complex route structures. However, an intelligent algorithm can be formulated by taking into account the historical trajectory of the vehicle and topological information of the road network (e.g., connectivity and orientation of links) to precisely identify the correct link on which a vehicle is traveling. Furthermore, an estimation of the vehicle location on the link can also be determined by taking into account all error sources associated with the positioning systems and digital map database. This is known as a map-matching algorithm. This chapter discusses the considerable momentum in research and development activities in map-matching, especially map data quality, methods and reliability issues surrounding map-matching algorithms. Future developments of map-matching algorithms and how such algorithms can tackle the positioning and navigation requirements of autonomous navigation are also discussed.

This chapter is organized as follows. First of all, an overview of spatial road map and map quality is provided. This is followed by a description of map-matching methods developed by researchers across the world. Map-matching integrity is then discussed. This chapter ends with a discussion on how a map-matching algorithm can aid in the navigation services of intelligent vehicles.

1 Spatial Road Map

A road network map database is a vital component for any Intelligent Transportation Systems (ITS) requiring spatial and temporal positioning and navigation data (Scott and Drane 1994). The road map is a graphical representation of the important spatial and topological information that a driver needs to negotiate during a trip, especially to a new location and hence it serves as the interface between the driver and the positioning and navigation technologies being used. A compact, informative road map is only the end product of a long and complex map creation process whose principles and procedures have been advanced over centuries (Drane and Rizos 1998). This section discusses some issues related to digital map errors and its quality.

1.1 Errors in Map Creation

The creation of a road network map involves a series of decisions on how features of the earth will be represented in an electronic map. Every such decision introduces a potential error in

the final map. According to the National Research Council (NRC 2002), the steps in the process include: map scale, level of generalization, projection, datum, and coordinate system.

Map scale can be defined as the ratio of distance on a map over the corresponding distance on the ground, represented as 1:M where M is the scale denominator. Map scale is an issue because as scale becomes larger the amount of detail that can be presented in a map is also increased. The ability to measure the length of linear features on the ground (road centerline), the position of point features (junctions and roundabouts), and the areas of polygons with a high level of accuracy are also increased. Map scales are usually divided into three categories, namely, large, medium, and small. A large-scale map extends over the range 1:1 to 1:24,000. An example of a large-scale map is a UK road network map (Land-line.Plus) at a scale of 1:1,250 developed by the UK Ordnance Survey. Medium-scale maps range from 1:24,000 to 1:100,000. Anything smaller than 1:100,000 are considered small scale. Maps for ITS applications should have a larger scale within urban areas compared to rural areas due to the typically higher density of road network features.

Whenever the surface of the earth is represented on a map, the features on the earth are generalized to some degree (NRC 2002). For example, roads are represented as a single “centerline” and curvatures are represented as piecewise linear lines (for a gentle curve) or as a polyline (for a sharp curve). This generalization reduces the accuracy of the representation of the features on the ground and can introduce significant bias.

Map projection is a technique for presenting data from a curved surface on a flat surface, for example, computer screen, sheet of paper, etc. There are many methods for map projection. The most common one is the Universal Transverse Mercator (UTM) projection. Changing a map projection implies simultaneously changing the relationships of area, shape, and direction on a map. Each of these factors can introduce error into the representation of a point, line, and area on a map.

The fourth choice that may introduce further error in the map is the selection of horizontal and vertical datum. A datum is defined as a reference of a quantity for calculation of other quantities (Czeriak and Reilly 1998). For a relatively small-scale map, a datum can have a serious effect on the placement of GPS positions onto a digital base map (NRC 2002).

The other potential error source is the coordinate system used. Most GPS receivers give positioning data in the WGS84 (WGS 84 is the 1984 revision of the World Geodetic System used for global datum system for horizontal datum.) (i.e., Latitude and Longitude) coordinate system. However, the users normally exploit their local coordinate systems. For example, the UK road network data is in the British Grid coordinate system (i.e., Easting and Northing). Therefore, transformation, conversion, and projection are essential to bring both systems to the same coordinate system. The process may also introduce error (NRC 2002).

1.2 Errors in Digitization

Digitization of a road map typically involves recording the spatial road data (point, line, and polygon) along with their associated labels and attributes digitally into a GIS environment. As discussed, there are numerous errors associated with a paper map that will be

automatically transmitted to the digital map format. Moreover, the digitizing process (entering spatial data using a digitizing tablet and puck) has its own inherent error that most likely will increase the error of the digital map. The Census Bureau found that a misalignment of the digitizing puck by 0.05 inch from a line on the paper map changed a point's position by 416 ft on the new digital map during the digitization of the US Geological Survey (USGS) paper topographic maps of source scale 1:1,000,000 (NRC 2002). This type of error is common when digitizing a paper map. Moreover, a variation in a room's temperature and humidity may shrink or expand the unstable base to which the paper map is attached, and hence introduce error.

1.3 Numerical Data Related to a Spatial Road Map

Topological features on the road network include both nodes and links. Curved roads are normally represented as polylines and straight roads are represented as lines in a digital map. In other words, arcs (roads) without *shape points* are referred to as lines and arcs with *shape points* are referred to as polylines. Each polyline consists of a series of lines depending on the number of *shape points* within the arc. Each arc is assumed to be piece-wise linear lines.

For simplicity, each *shape point* is assumed to be a node. Each line and node is associated with a set of identifying attributes. The attributes of the node are the *x*-coordinate and *y*-coordinate that identify the spatial position of the node (► Table 14.1a). A node with the same *x*-coordinate and *y*-coordinate has the same identification number. The attributes of each line (i.e., link) are determined from the nodes within the line (i.e., start node and end node) as in ► Table 14.1b. Therefore, connectivity information among links at a junction can be derived from these two geographic data files and can then be used as an important input to a map-matching algorithm. Additional node attributes such as grade separation and turn restriction and link attributes such as link length, the direction of traffic flow in a two-way road and the number of lanes could aid in the process of map-matching.

2 Digital Map Quality

Digital map data is usually based on a single-line-road-network representing the center-line of the road. Road attributes such as width, number of lanes, turn restrictions at junctions, and roadway classification (e.g., one-way or two-way road) normally do not exist in the map data. Therefore, the accuracy and uncertainty of digital road network data is a critical issue if the data are used for land vehicle navigation. One must be aware of the following concerns regarding the quality of road network data (Quddus 2006):

- The features (e.g., roundabouts, junctions, medians, curves) of the real world that have been omitted or simplified in the road map. This is usually known as topological error.

■ Table 14.1

Numerical format of (a) node and (b) link data

(a)		
Node ID	Easting	Northing
1	525841	178856
2	525488	178626
3	525463	178656
4	525432	178703
5	525498	178617
6	525399	178772
7	525933	178863
8	525320	178917
(b)		
Link ID	Start_node	End_node
1	1	7
2	2	3
3	3	4
4	4	6
5	5	2
6	6	13
7	7	10
8	8	167

- The accuracy of the classification (e.g., junction or roundabout) of those features.
- Data currency, that is, how currently the map was created.
- The displacement of a map feature (e.g., road centerline, specific junction) from its actual location in the road. This is generally known as geometric error.

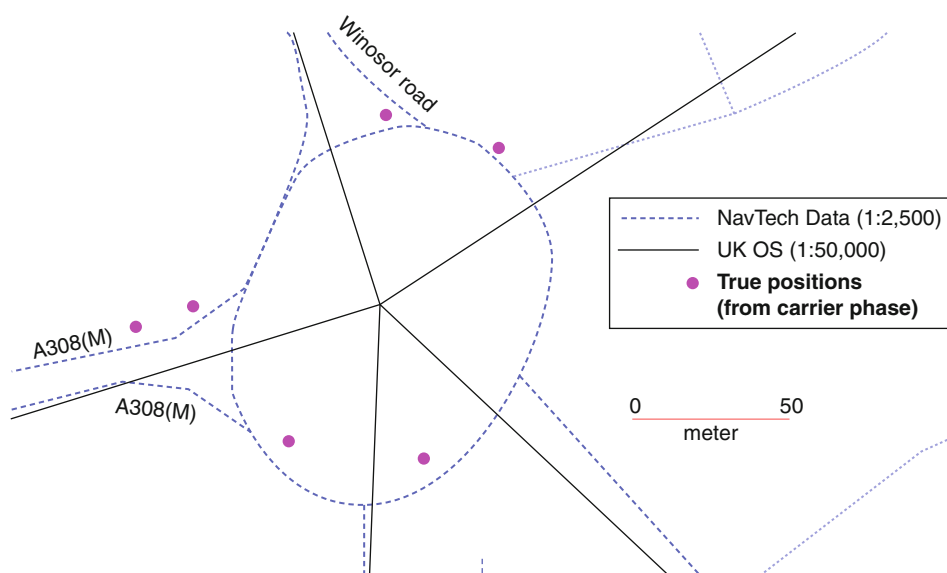
Both geometric and topological errors of map data may introduce significant horizontal errors in land vehicle positioning and navigation. While the geometric error can be corrected with suitable hardware, software, and algorithms, the topological error cannot be easily corrected (Goodwin and Lau 1993). The accuracy (2D) of a digital map is defined as the closeness of a measurement or estimate to a true value. A rigorous statement of accuracy includes statistical measures of uncertainty and variation. Accuracy is generally represented by the standard deviation of errors (difference between measurements on the map and the corresponding true values). The true values of map attributes or labels can be obtained from an independent source of higher accuracy measurement such as GPS carrier phase observables, higher resolution satellite imagery, aerial photography, or a ground visit. If e^1, e^2, \dots, e^n represent a series of differences between measurements on a map and their true values, the accuracy of this map is given by σ_{map}^2 , where

$$\sigma_{map}^2 = \frac{\sum_{i=1}^n (e^i - \mu)^2}{n} \quad (14.1)$$

in which $\mu = \frac{\sum_{i=1}^n e^i}{n}$ and n is the number of measurements.

There are also empirical methods available to estimate map accuracy. For instance, the University of Texas at Austin-Department of Geography (UTADG) developed a method to compute map accuracy from knowledge of the errors introduced by different sources. This method calculates an estimate of overall accuracy by summing the squares of specified components of the map and taking the square root of the sum. For example, assuming the scale of a map is 1:2,500, the estimated error due to the source document is 2.5 m (1 mm*2,500), map registration is 1.25 m (0.5 mm*2,500), and digitizing is 0.5 m (0.2 mm*2,500), the total error of a map of scale 1:2,500 is 2.84 m. Following this methodology, the total estimated error of the map scale 1:25,000 is 28.39 m.

► Figure 14.1 shows the graphical comparison of two digital road maps (road center-line) with map scales 1:2,500 (from NavTech) and 1:50,000 (from the UK Ordnance Survey). The true vehicle positions obtained from higher accuracy GPS carrier phase observables are denoted by the dot symbols. The NavTech map data and the true vehicle positioning data agree reasonably well and suggest that the section of the road map is a roundabout. However, the UK Ordnance Survey map data indicate that the section of the map is rather a five-legged junction. If 1:50,000 scale map is used in vehicle navigation, the horizontal positioning error may be increased significantly.



■ Fig. 14.1

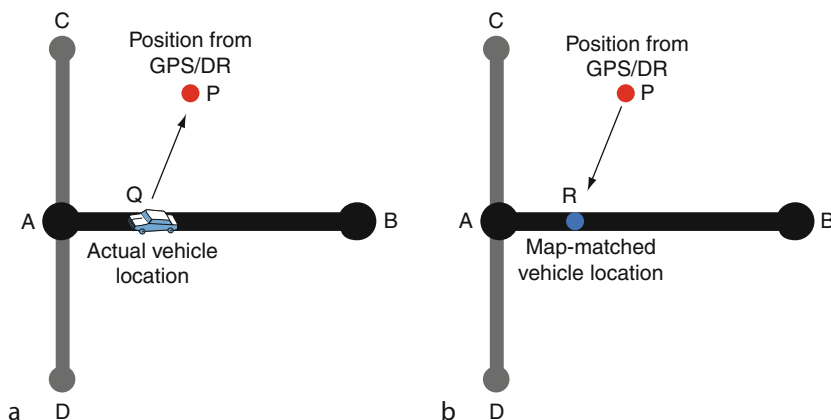
A graphical comparison of road network data from two different sources

It should be noted that the accuracy of a digital map determined by (14.1) may not always indicate a better quality road map. This is due to the fact that the authority responsible for managing road infrastructure is continuously changing road configurations (e.g., from a one-way street to a two-way street and vice-versa) or developing new roads to enhance overall traffic operations. Vendors who produce the digital maps are not always aware of such modifications and therefore, digital maps may characterize “missing links” and this may affect the performance of a navigation system. It is therefore vital that vendors update their digital maps periodically so as to minimize the impact of digital map quality on the navigation services.

3 Methods of Map-Matching

The purpose of map-matching is to integrate the positioning data with the spatial road network data, to identify the actual link on which a vehicle is traveling and to determine the vehicle location on that link. Assume that P represents the position fix obtained from the navigation system when the actual location of the vehicle is at Q (Fig. 14.2a). P deviates from the road centerline due to errors in the map and the navigation data. A map-matching algorithm identifies the correct link (AB) for the position fix P and snaps back the position fix onto link AB (Fig. 14.2b).

Most of the existing map-matching algorithms are based on the theories outlined in Zhao (1997) with enhancements proposed by other researchers including White et al. (2000), Greenfeld (2002), Quddus et al. (2003), Quddus (2006), etc. Procedures for map-matching vary from those using simple search techniques (Kim et al. 1996), to those using more complex mathematical techniques such as a Kalman Filter (Kim et al. 2000). Approaches for map-matching algorithms found in the literature can be categorized into three groups (Quddus 2006): geometric, topological, and advanced. The following sections briefly describe these algorithms.



■ Fig. 14.2
Map-matching concept

3.1 Geometric Analysis

A geometric map-matching algorithm makes use of the geometric information of the digital road network by considering only the shape of the links (Greenfeld 2002). It does not consider the way links are connected to each other. The most commonly used geometric map-matching algorithm is a simple search algorithm. In this approach, each of the positioning fixes matches to the closest “node” or “shape point” of a road segment. This is known as point-to-point matching (Bernstein and Kornhauser 1996). A number of data structures and algorithms exist in the literature to select the closest node or shape point of a road segment from a given point (e.g., Bentley and Mauer 1980; Fuchs et al. 1980). This approach is both easy to implement and very fast. However, it is very sensitive to the way in which the network was digitized and hence has many problems in practice. That is, other things being equal, arcs with more shape points are more likely to be properly matched. In a straight arc with two end nodes, all positioning points above the arc match only with the end nodes of the arc. The details of this point-to-point map-matching method are provided in Quddus (2006).

Another geometric map-matching approach is point-to-curve matching (Bernstein and Kornhauser 1998; White et al. 2000). In this approach, the position fix obtained from the navigation system is matched onto the closest curve in the network. Each of the curves comprises line segments, which are piecewise linear. Distance is calculated from the position fix to each of the line segments. The line segment which gives the smallest distance is selected as the one on which the vehicle is apparently traveling. Although this approach gives better results than point-to-point matching, it does have several shortcomings that make it inappropriate in practice. For example, it gives very unstable results in urban networks due to the high road density. Moreover, the closest link may not always be the correct link (Quddus 2006).

The other geometric approach is to compare the vehicle’s trajectory against known roads. This is also known as curve-to-curve matching (Bernstein and Kornhauser 1996; White et al. 2000). This approach firstly identifies the candidate nodes using point-to-point matching. Then, given a candidate node, it constructs piecewise linear curves from the set of paths that originates from that node. Secondly, it constructs piecewise linear curves using the vehicle’s trajectory, that is, the positioning points, and determines the distance between this curve and the curve corresponding to the road network. The road arc, which is closest to the curve formed from positioning points, is taken as the one on which the vehicle is apparently traveling. This approach is quite sensitive to outliers and depends on point-to-point matching, sometimes giving unexpected erroneous results.

Enhancements of geometric map-matching algorithms are provided in Taylor et al. (2001), Bouju et al. (2002), Phuyal (2002) and Srinivasan et al. (2003).

3.2 Topological Analysis

Further improvements of geometric map-matching algorithms were achieved by considering additional information in correct road segment identification process. Topological

map-matching uses additional information including historical/past matching information, vehicle speed, vehicle turn restriction information, and topological information of the spatial road network (e.g., link connectivity) in addition to road geometry (Greenfeld 2002; Quddus 2006; Xu et al. 2007; Pink and Hummel 2008; Velaga et al. 2009). This additional information enables topological map-matching algorithms to outperform geometric approaches. The basic advantages of a topological map-matching algorithm over geometric algorithms are:

1. Positioning points are not treated individually; historical/past matching information is used.
2. The correct road segment identification process is more logical.
3. Topological information of the spatial road network (e.g., link connectivity) and additional information, such as vehicle speed and heading are used in the correct link identification process.

Generally, any map-matching process consists of three key stages: (a) initial map-matching, (b) matching on a link, and (c) map-matching at a junction. The aim of the initial map-matching process is to identify the correct road segment for the first positioning point. After snapping the first positioning point to a selected link, the algorithm checks whether the vehicle is traveling on the same link which was selected for the first positioning point or the vehicle is near a downstream junction. To examine how far the vehicle is from the downstream junction, information like the distance from the previously map-matched vehicle position to the downstream junction, distance traveled by the vehicle in the last time interval, and change in vehicle movement direction (i.e., vehicle heading) with respect to the previously selected link direction are used. In case of the second stage, *matching on a link*, the algorithm directly snaps the vehicle to the previously selected road segment for the last positioning point. If the vehicle reaches downstream junction, the map-matching algorithm needs to identify the correct road segment on which the vehicle is traveling. For *initial map-matching* and *map-matching at a junction* the algorithm follows three successive steps: (1) identification of a set of candidate links around the positioning point, (2) selection of correct link from the candidate links, and (3) determination of vehicle position on the selected road link. The typical part of any map-matching algorithm is to identify the correct road segment from a set of candidate road segments. Map-matching performance also depends on a set of candidate link identification process. If the correct road segment on which a vehicle is traveling is not included in the candidate link set the result may lead to a wrong map-matching. Few existing map-matching algorithms also carry out consistency checks before finalizing the selection of the correct link among the candidate links. These consistency checks are to avoid any possible mismatching. The phenomenon of identification of wrong road segment from a set of candidate road segments is known as mismatching.

Various topological map-matching algorithms have been developed to support location-based intelligent transport services. Different authors have used the topological information at various levels. For instance, White et al. (2000) and Li and Fu (2003) used topological information to identify a set of candidate links for each positioning point. Srinivasan et al. (2003) and Blazquez and Vonderohe (2005) used topological

information to check the map-matched point after geometric (point-to-curve) matching. Greenfeld (2002), Quddus et al. (2003), Taghipour et al. (2008) and Velaga et al. (2009) introduced weight-based algorithms to identify the correct road segment among the candidate segments for real-time map-matching problem.

The earlier topological map-matching algorithms recognize the nearest node or shape point for a positioning point and the links connected to that node or shape point are considered as candidate links for that positioning point. White et al. (2000) and Li and Fu (2003) used vehicle heading information (i.e., vehicle movement direction with respect to the North) to identify a set of candidate links for each position point. If the vehicle heading is not in line with the road segment direction, then the road segment is discarded from the set of candidate segments. If the algorithm has confidence in the previous map-matched position it uses the topology (road connectivity) of network for current positioning point map-matching. That means the algorithm considers the road segments that are connected to the previous map-matched road link to identify the current road segment. Few recently developed algorithms create an error circle around the positioning point. The radius of the error bubble is primarily based on the quality of positioning data (i.e., variance and covariance of easting and northing) at that instant (for that positioning point). All the links that are either inside the error bubble or touching the error bubble are considered as the candidate links for that positioning point.

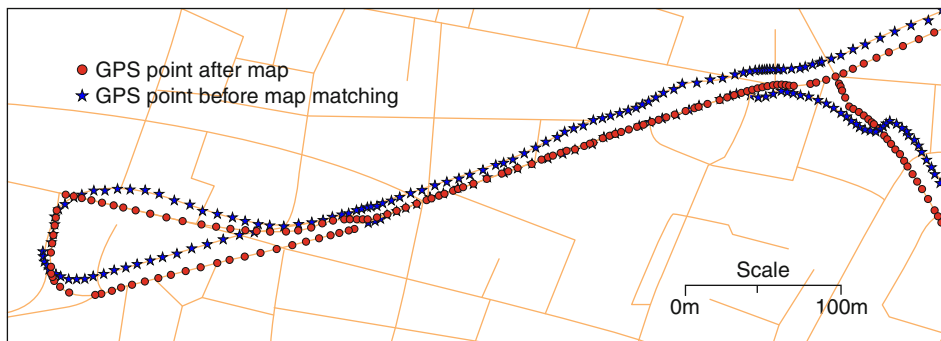
A weight-based topological map-matching algorithm assigns weights for all candidate links based on different criteria such as the similarity in vehicle movement direction and link direction and the nearness of the positioning point to a link (Greenfeld 2002; Quddus et al. 2003). Among the candidate links, greater weight should be given to a link that is in-line with the vehicle movement direction. Similarly, if a link is nearer to the positioning point, then this link should be given more weight than a link which is further away. Topological information of the spatial road network (e.g., link connectivity and turn restriction) are also used as weighting parameters in selecting the correct link from a set of candidate links (Velaga et al. 2009). For instance, a road link should be given more weight if it is directly connected to the previously traveled road link and the vehicle is legally allowed to make a turn onto the link. However, link connectivity and turn restriction information cannot be useful for the initial map-matching (i.e., map-matching for the first positioning point). This is simply because the previously traveled road segment for the first positioning point is not known. The candidate link with the highest total weighting score is selected as the correct link at the junction. The perpendicular projection of the positioning point onto the correct link provides the vehicle location on the road segment. The total weighting score is given as follows (see Velaga et al. 2009 for details).

$$TWS_j = W_h + W_p + W_c + W_t \quad (14.2)$$

where,

TWS_j is the total weighting score;

W_h is the weighting score for heading, which represents the similarity in vehicle movement direction and link direction;



■ Fig. 14.3
Map-matching output on a GIS map

W_p is the weighting score for proximity (nearness of a positioning point to a link);
 W_c is the weighting score for link connectivity to the previously traveled road link and;
 W_t is the weighting score for turn restriction.

This type of map-matching algorithm is very popular due to its simplicity and speediness in identifying the correct links. A recent topological map-matching algorithm developed by Velaga et al. (2010a) succeeded 97.8% of the time in correct link identification with a horizontal accuracy of 9.1 m ($\mu + 2\sigma$) in urban areas in Nottingham, UK. An example of map-matching results in a dense urban area in London is shown in Fig. 14.3. The raw positioning points are represented with star symbols and map-matched points are shown with round symbols. The performance of a topological map-matching algorithm is better than that of a geometric algorithm. Topological algorithm may also fail in identifying the correct road segments, especially at “Y” junctions, roundabouts, and dense urban areas.

3.3 Analysis with Advanced Techniques

In dense urban areas, where a road network is very complex and vehicle positioning data from GPS or other positioning systems suffer from signal masking and multipath errors due to tall buildings, tunnels, and narrow roads, it is very difficult to select the correct road segments on which a vehicle is traveling. In urban areas, raw GPS positioning data often associates with high errors and the number of candidate links for a positioning point are also high. Consequently, the topological map-matching approaches which use similarity in vehicle movement direction and link direction, nearness of positioning point to a link and link connectivity may not be adequate to precisely identify the road on which the vehicle is traveling. Therefore, more advanced methods and artificial intelligent techniques need to be used in the map-matching process.

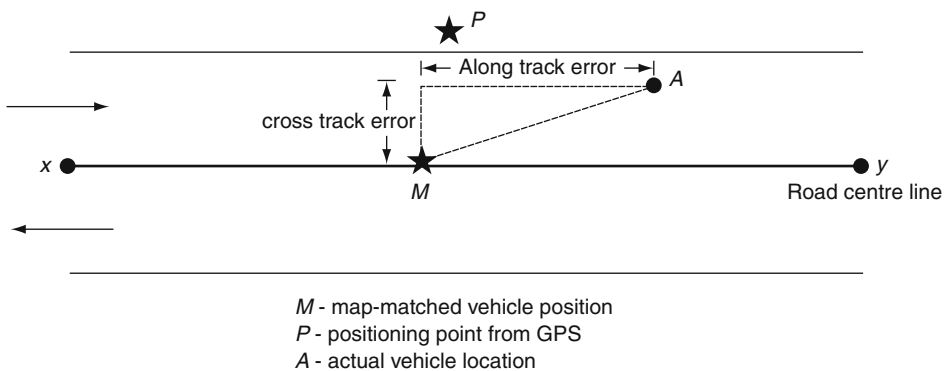
As mentioned, the most critical part of any map-matching process is to detect the correct road segment among the candidate road segments. In most algorithms, the advanced techniques are used in the correct link identification process. These advanced algorithms use more refined concepts such as: Extended Kalman Filter (EKF) (e.g., Krakiwsky 1988; Tanaka et al. 1990; Jo et al. 1996; Kim et al. 2000; Torriti and Guesalaga 2008), Bayesian interference (e.g., Pyo et al. 2001; Smaili et al. 2008), belief theory (e.g., Yang et al. 2003; Najjar and Bonnifait 2005; Nassreddine et al. 2008), fuzzy logic (e.g., Zhao 1997; Kim and Kim 2001; Syed and Cannon 2004; Quddus et al. 2006a; Su et al. 2008; Zhang and Gao 2008) and artificial neural networks (e.g., Ding et al. 2007; Su et al. 2008). Among these three map-matching approaches (geometric, topological, and advanced) an advanced map-matching algorithm, that uses more refined concepts, offers the highest performance, both in terms of link identification and location determination. However, these advanced map-matching algorithms require more input data and are relatively slow and difficult to implement compared to a topological map-matching algorithm (Velaga et al. 2009).

Any map-matching algorithm requires a range of data input such as positioning data, topological features of spatial road network data (White et al. 2000). The positioning data includes GPS easting and northing (i.e., X and Y coordinates), vehicle heading (i.e., vehicle movement direction with respect to the North direction), and vehicle speed data in m/s. Digital map data includes node data and link data. The node data consists of node number, node easting, and northing (i.e., X and Y coordinates). The link data consists of link ID, start node, and end node of each link (see ► Table 14.1). An example of such node and link data is shown in ► Table 14.1. Other input data might include connectivity among road links, legal turn restriction information at junctions, legal speed limit, road classification (e.g., one way or two way), quality of position solution (e.g., Horizontal Dilution of Precision, HDOP), position of raw GPS points relative to a candidate link and digital road map errors. A number of new input variables (at no extra cost) are generally included in advanced algorithms to precisely estimate the correct road on which a vehicle is traveling and the vehicle location on that selected road. For example, a fuzzy logic-based map-matching algorithm developed by Quddus et al. (2006a) considered the following extra information in addition to the basic vehicle positioning data and digital map data: (1) the speed of the vehicle, (2) the connectivity among road links, (3) the quality of position solution, for example, HDOP, and (4) the position of a fix relative to a candidate link.

The basic process in most of the advanced map-matching algorithms is quite similar to the sequential steps explained in the topological map patching process. That is (1) candidate link identification (2) correct link identification (3) vehicle location identification on the selected link and followed by consistency checks. For example, a fuzzy logic-based map-matching algorithm developed by Quddus et al. (2006a) divided the map-matching into two basic stages: The processes are: (1) the *initial map-matching process* and (2) the *subsequent map-matching process*. Both the initial and subsequent processes follow the above-mentioned sequential steps. An error ellipse is derived based on the quality of positioning fix in easting and northing. All the links inside the error ellipse are considered as candidate links for that positioning point. Two Sugeno's Fuzzy Inference Systems were

developed for *initial map-matching* and *subsequent map-matching*. For the first positioning point, in addition to the basic variables such as heading difference (absolute difference between the direction of the vehicle and the direction of the link) and perpendicular distance from the position fix to the link, speed of the vehicle and quality of positioning point are also considered. Therefore, the state input variables of the Fuzzy Inference System are: (1) the speed of the vehicle, v (m/s), (2) the heading error (degrees), (3) the perpendicular distance (in meters), and (4) the HDOP Quddus et al. (2006a). Six fuzzy rules are considered in the analysis and a weighted average method is used to obtain a crisp output, which gives the likelihood associated with a link. The Fuzzy Inference System is applied to all the candidate links and the link with the highest likelihood is taken as the correct link among the candidate links. The *subsequent map-matching process* is almost similar to the initial matching process, but only difference is three additional input variables and thirteen rules.

Though these advanced techniques are used in the correct link identification process, the two-dimensional horizontal accuracy is also enhanced. Because both link identification and positioning accuracy are interlinked. If a vehicle is assigned to a wrong road segment, the positioning accuracy will be low and vice versa. Often the horizontal accuracy is represented in two components: (1) along track error and (2) cross track error. Along track error represents the positioning error of map-matched point along the road and cross track error across the road. In Fig. 14.4, raw positioning point P is assigned to the road central line (x, y) at point M . Point A is the actual vehicle location for that positioning point. The actual (true) position can be obtained from an independent source of high accuracy measurement such as GPS carrier phase observables (i.e., observations with accurate GPS) and a high-grade inertial navigation system. Here, the distance between actual vehicle location (i.e., point A) and the map-matched vehicle location (i.e., point M) is the horizontal error, and its corresponding components are the along track error and cross track error.



■ Fig. 14.4
Map-matching error

The fuzzy logic-based map-matching algorithm, which takes input from GPS/Dead reckoning and high-quality digital map (scale 1:1,250), developed by Quddus et al. (2006a) is capable of identifying 99.2% of the links correctly with the horizontal positioning accuracy of 5.5 m (95% confidence interval) for suburban road network near London. The corresponding along track error and cross track error were found to be 4.2 m and 3.2 m, respectively. A map-matching algorithm by Bonnifait et al. (2009) is succeeded in identifying correct road segments 99.7% of the time; However, their study was conducted in a suburban area and did not measure the horizontal accuracy due to lack of accurate (true) positioning information. In addition to performance in terms of correct link identification and positioning accuracy, computational speed of the map-matching algorithms is also important because map-matching needs to be performed in real time.

The accuracy of a map-matching algorithm not only depends on the methods (i.e., geometric, topological, and probabilistic) used in the map-matching algorithm and the quality of raw GPS outputs but also rely on the quality of digital maps. It has been identified that there are considerable effects of the quality of spatial road network data on the performance of map-matching algorithms. For example, Quddus et al. (2008) examined the performance of three different map-matching algorithms using a vehicle positioning data obtained from two different systems (GPS and GPS integrated with DR) and three types of spatial road network maps of different map scales (1:1,250, 1:2,500, and 1:50,000). The performance of a geometric map-matching algorithm, for instance, in correct link identification, using data from an integrated GPS/DR system, was found to be 88.7%, 87.5%, and 75.5% with map scales 1:1,250, 1:2,500, and 1:50,000, respectively, suggesting that the performance of the algorithm is better with the good quality base map. This is also true for other map-matching approaches.

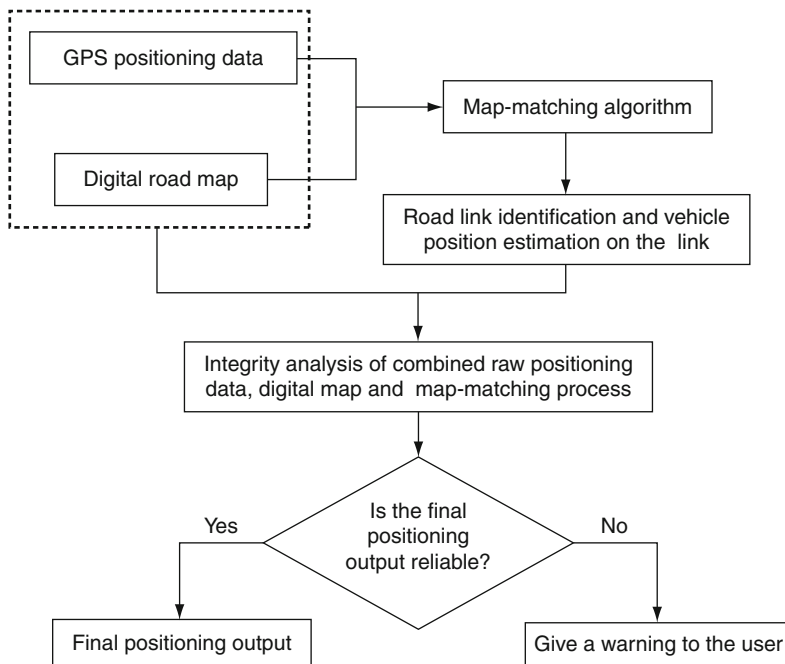
4 Integrity and Reliability of Map-Matching

Although the most sophisticated and advanced techniques are used in the map-matching process vehicle mismatching (i.e., snapping vehicle to a wrong road) is sometimes inevitable. Any error associated with either the raw positioning points, the digital map used, or the map-matching process employed can lead to a mismatch. The raw positioning data from a navigation system contains errors due to satellite orbit and clock bias, atmospheric (ionosphere and troposphere) effects, receiver measurement error, and multipath error (Kaplan and Hegarty 2006). As discussed, GIS-based road maps include errors which can be geometric (e.g., displacement and rotation of map features) and/or topological (e.g., missing road features) (Goodwin and Lau 1993; Kim et al. 2000). Even when raw positioning data and map quality are good, map-matching techniques sometimes fail to identify the correct road segment especially at roundabouts, level-crossings, Y junctions, and dense urban roads and parallel roads (White et al. 2000; Quddus et al. 2007).

Users should be notified when the system performance is not reliable. This is vital for safety critical applications such as emergency vehicle management, vehicle collision avoidance, and liability critical applications such as electronic toll collection system and

distance-based pay-as-you-drive road pricing. For example, the wrong vehicle location identification due to errors associated with a navigation system used in an emergency vehicle routing service delays an ambulance arrival at the accident site; this may lead to loss of life. If the user is informed when system performance is not reliable then the user will not blindly depend on the navigation system.

The concept of integrity monitoring has been successfully applied to air transport navigation systems. Very little research has been devoted to land vehicle navigation system integrity monitoring. The primary difference is that in addition to the errors associated with the space segment (GPS satellite system) data processing chain, one needs to consider the errors associated with the digital map and the map-matching process when monitoring the integrity of a land vehicle navigation system. Research on land vehicle integrity monitoring has concentrated on either the integrity of raw positioning data obtained from GNSS (Philipp and Zunker 2005; Feng and Ochieng 2007; Lee et al. 2007) or the integrity of the map-matching process (Yu et al. 2006; Quddus et al. 2006b; Jabbour et al. 2008). Velaga et al. (2010b) attempted to consider these error sources together along with complexity of the road network (i.e., operational environment). Figure 14.5 provides a basic integrity process. Considering both sources of error concurrently, in identifying the goodness (trustability) of final positioning point, lead to a better outcome. Moreover, taking the complexity of the road network (i.e., operational environment) into account may further improve the integrity process. This is because, although the raw positioning



■ Fig. 14.5
Integrity process

points from GPS contain some errors, a good map-matching algorithm can identify the correct road segment if the road network in which the vehicle is traveling is not complex.

Integrity monitoring of a raw GPS positioning point can be categorized into system, network, and user-level monitoring (Feng and Ochieng 2007). System-level monitoring can provide integrity information by considering satellite orbit and clock bias. An example of system-level integrity is Galileo. Network-level integrity monitoring can be further classified into: Satellite-Based Augmentation System (SBAS) and Ground-Based Augmentations System (GBAS) (Bhatti et al. 2006). The SBAS is an overlay system to enhance the accuracy of GPS by using additional satellites. The SBAS considers satellite orbit and clock corrections, geometry of satellites and user, and ionosphere error, but not the tropospheric effects, receiver measurement error and multipath error (Feng and Ochieng 2007). An example of SBAS is European Geostationary Navigation Overlay Service (EGNOS), operated by the European Space Agency. In case of GBAS, ground survey stations monitor the health of satellite and transmit correction to the user receivers via very high frequency transmitters. The GBAS considers troposphere correction in addition to the other errors covered in SBAS. An example of GBAS is the USA's Local Area Augmentation System (LAAS). Neither system-level nor network-level integrity monitoring cover multipath error that generally occurs in urban canyons.

Integrity monitoring, at the individual user level which considers receiver measurement errors and multipath error along with all the other errors discussed above is referred to as the User-Level Integrity Monitoring (ULIM) process. A ULIM with a stand-alone GNSS is generally called Receiver Autonomous Integrity Monitoring (RAIM). Integrity monitoring combining GNSS and other sensors (e.g., dead reckoning system or inertial navigation system) is commonly known as User Autonomous Integrity Monitoring (UAIM) (Ochieng et al. 2007). The UAIM is a special case of RAIM. The UAIM considers sensor (odometer and gyroscope) errors along with GPS receiver measurement errors and multipath error. Basic methods that are used to calculate user-level integrity monitoring of raw positioning points are: (1) Range comparison method (2) Least square residual method (3) Weighted least square residual method (4) Kalman filter method.

Most of the existing map-matching algorithms carry out consistency checks after selecting correct road segment from a set of candidate segments and before providing final map-matched positioning point. Example of such checks are comparing distance between previously map-matched positioning point and current map-matched point with the distance calculated based on speed of the vehicle in the last time fix, comparing the difference between a raw position point and a map-matched position on the link with the estimated error in raw positioning point. However, these consistency checks are useful to avoid any possible blunder errors in the final map-matched positioning output; but these checks cannot provide the confidence level of the map-matched output. To provide the trustability level of map-matching process few authors measure the integrity of map-matching algorithms. For example, Quddus et al. (2006b) developed an integrity method for map-matching algorithms using *sugeno fuzzy inference* system; Jabbour et al. (2008) proposed a map-matching integrity method using a multihypothesis technique; Lee et al. (2007) developed an integrity method using a Monte Carlo technique for land vehicle navigation.

Velaga et al. (2010b) developed a map-aided integrity monitoring process for land vehicle navigation. This integrity method considered all the error sources associated with raw positioning point, map-matching process, and digital map errors together along with complexity of the road network (i.e., operational environment). In this method, firstly, a weighted least squares residual method is used to identify the fault in raw positioning process. If no fault is detected in the raw positioning point, then the map-matching process and its integrity measurement are carried out. This is because, although there is no error with the raw GPS point, the map-matching process may fail to identify the correct link (from the set of candidate links) due to the complexity of road network, errors in GIS digital map, or errors in the map-matching process. If any fault is detected in raw positioning process, rather immediately giving a warning to the user, the integrity method identifies the operational environment in which the vehicle is traveling. Although the raw positioning points from the GPS contains some errors, a good map-matching algorithm can identify the correct road segment if the road network on which a vehicle is traveling is simple. Two different fuzzy inference systems are designed for fault-free raw positioning point and faulty raw positioning point. The output from fuzzy system, provides an integrity score (ranges from “0” to “100”), which gives the confidence level of final map-matched positioning output. The value “0” represents the most un-trustable user position and “100” represents the most trustable positioning point.

The criterion that is normally employed to evaluate the performance of an integrity method is Overall Correct Detection Rate (OCDR) (Quddus et al. 2006b; Jabbour et al. 2008). The OCDR refers to percentage of time the system can provide valid warnings to users. Further, invalid warnings can be either *missed detections* or *false alarms*. The missed detection (MD) suggests that although there is a mistake in the final positioning output of the navigation system, the integrity method could not identify it; and the false alarm (FA) indicates that although there is no error in the final positioning output obtained from the vehicle navigation module the system gives an alert to users. So the performance of an integrity method could be measured with respect to Missed Detection Rate (MDR), False Alarm Rate (FAR) and Overall Correct Detection Rate (OCDR) (Quddus et al. 2006b; Jabbour et al. 2008). As suggested by Quddus et al. (2006b) and Jabbour et al. (2008) the overall correct detection rate (OCDR), which is derived with respect to false alarm rate (FAR) and missed detection rate (MDR), can be written as:

$$OCDR = 1 - (FAR + MDR) \quad (14.3)$$

$$FAR = \frac{f}{o} \quad (14.4)$$

$$MDR = \frac{m}{o} \quad (14.5)$$

where,

f is the total false alarms,

m is the total missed detections and

o is the total observations.

The integrity method developed by Velaga et al. (2010b) provides OCDR, FAR, and MDR 0.9817, 0.0084, and 0.0099, respectively, when tested with 2,838 positioning points obtained from urban areas of Nottingham, UK. That means, the integrity method gave correct warnings 98.2% of the time. To decide the total number of false alarms (f) and the number of missed detections (m), reference (true) vehicle positioning is required. Here, the positioning points obtained from GPS carrier-phase observations integrated with a high-grade Inertial Navigation System (INS) were used as reference positioning points. Further research may be required in integrity of land vehicle navigation system to support a few safety critical in-vehicle applications such as collision avoidance systems and autonomous driving.

5 Intelligent Vehicles and Map-Marching Algorithms

The twentieth century has witnessed the automobile development from a simple horseless carrier to one of the most technologically advanced mass market commodities available (Young et al. 2007), and recent efforts by the automobile industry and academic researchers are being devoted to automate most aspects of vehicle operations in which increased safety, efficiency, and comfort are cited as “the driving force” behind vehicle automation. One of the congressional mandates states that, “a third of U.S. military ground vehicles be unmanned by 2015” (Urmson et al. 2008, p. 426). Recent advances in sensors, computing, communication technologies, and 3-D digital maps create the possibility of developing autonomous navigation of *intelligent agents* (not only intelligent vehicles – iVehicles – but also robots, intelligent wheelchairs) a reality. For instance, intensive research and development efforts have been dedicated to the possibility of making intelligent vehicles (iVehicles) that have the capability of fully autonomous driving in contrast to currently available assisted driving. Grand projects include: (1) the 800 million (€) EC EUREKA Prometheus Project, (2) the DARPA Grand Challenge from the USA, and (3) the ARGO research project from Italy. Previous research identified a range of benefits that are expected from iVehicles: reducing fuel consumption, eco-driving, enhancing safety (i.e., *zero fatality*, reducing severity of occupant injuries), helping ageing society, and navigating miles of off-road terrain for search and rescue services (i.e., military applications).

One of the key components of such an *intelligent vehicle* is the localization and navigation module that estimates road geometry or localizing the agent relative to the road with known geometry. Researchers are facing major challenges while targeting “zero margin of error” in autonomous navigation for *intelligent vehicles*, especially in an organically changing high dynamic network (i.e., urban canyons with high traffic volume). This is primarily due to the lack of accurate 3-D spatial data (*geo-referenced digital images*) of surrounding environment which can be used in the process of map-matching algorithm for high precision localization. Existing map-matching techniques discussed above are suitable for matching 2-D positioning and navigation data with 2-D spatial road network data. For the case of autonomous navigation of intelligent vehicles, features in real-time images (3-D) captured by 3-D laser scanners need to be matched with the corresponding

features in the 3-D base map of surrounding environment (i.e., *geo-referenced digital colored images*) to enhance localization. The accuracy and efficacy can further be enhanced by camera imagery. Another approach would be to align a vehicle-based 3-D local map to a satellite-based 3-D global map, if available from a satellite-based laser altimeter. The very latest laser scanning technology coupled with a high precision navigation system can be used to develop a 3-D LIDARs by scanning roads, buildings, and trees from a moving vehicle. Such a model can then be employed as a base map in the map-matching process. A new technique needs to be developed to match real-time images with the 3-D base map. One of the main tasks would be to develop a series of methods to determine the location of intelligent agents based on the captured real-time images and compare them with the 3-D static map using a range of *feature-based approaches* such as spin images and Multi-frame Odometry-compensated Global Alignment (MOGA). Various artificial intelligence techniques can be employed to enhance images and to develop new map-matching algorithms (feature detection, matching, and pose refinement). A good uncertainty model is also required to correctly assess the quality of individual features and local-global matches.

Therefore, the fundamental challenge for efficient high precision localization and mapping of intelligent vehicles is to develop a navigation system that eliminates human involvement in determining the best course of travel. The system should be able to gather information from an external source (i.e., a traffic control center) and roadway signs (i.e., variable message signs, regulatory signs, warning signs, and information signs intended to assist the control computer in getting to their destinations) and then integrate them with various on-board sensors and map databases (3-D) for the purpose of determining dynamic real-time navigation. In order to realize such autonomous navigation systems, a number of further challenges must be overcome, the critical ones being extraction of features from the surrounding environment (i.e., road edges, road layouts, markings, and signs) for lane-based positioning and trajectory generation, integration of range measurements from Network-RTK (Real-time Kinematic), laser scanners, and Inertial Measurement Unit (IMU) to obtain continuous high accurate (centimeter-level) positioning data and development of dynamic route planning using multi-objective routing algorithms and map-matching algorithms. The overall integrity of the autonomous navigation systems should be achieved by proper treatment of all uncertainties associated with the measurements. In contrast to current navigation systems that are heavily dependent on on-board sensors, the future autonomous system will heavily be dependent on data from external sensors/data (i.e., Network-RTK, 3-D city model, and road signs) and therefore, the cost per unit would be relatively less. If successful, this will produce a step change in current practices of autonomous navigation and will lead to a cost-effective solution to autonomous driving.

6 Summary

The key underpinning process of in-vehicle navigation systems is map-matching. This chapter highlighted some of the recent developments of map-matching algorithms and documented relevant issues and challenges in map-matching. This includes: the quality of

digital map (including map updates), the methods used in the map-matching process, and integrity of map-matching. It has been shown that the performance of a map-matching algorithm not only relies on the methods employed in a map-matching process but also the quality of both positioning systems and digital road maps. The fundamental of developing an in-vehicle navigation system depends on the navigational and positioning requirements (in terms of accuracy, integrity, continuity, and availability) of the target applications. For instance, if an in-vehicle navigation system intends to support safety-of-life critical services (i.e., collision warning) then the selection of high accuracy and high integrity map-matching algorithms (along with a good digital map and high accuracy positioning sensors) should be the norm. In order to support semi- and full-autonomous navigation services, additional sensors (i.e., laser scanners, digital cameras) and data (i.e., a 3-D spatial road environment map) are needed. Therefore, future map-matching algorithms should be flexible to accommodate rapid developments in the quality and quantity of the sensor outputs and spatial network data. It would be interesting to see if such developments result in a reduction in the complexity of designing autonomous navigation systems.

References

- Bentley JL, Mauer HA (1980) Efficient worst-case data structures for range searching. *Acta Inf* 13:155–168
- Bernstein D, Kornhauser A (1996) An introduction to map-matching for personal navigation assistants. Available at <http://www.njtude.org/reports/mapmatchintro.pdf>. Accessed 19 June 2002
- Bernstein D, Kornhauser A (1998) Map matching for personal navigation assistants. In: *Proceedings of the 77th annual meeting of the transportation research board*, Washington D.C., 11–15 Jan 1998
- Bhatti U, Ochieng WY, Feng S (2006) Integrity of integrated GPS/low cost inertial system and failure modes: analysis and results – part 1. *GPS Solut* 11(3):173–182
- Blazquez CA, Vonderohe AP (2005) Simple map-matching algorithm applied to intelligent winter maintenance vehicle data. *Transp Res Rec* 1935:68–76
- Bonnifait Ph, Laneurit J, Fouque C, Dherbomez G (2009) Multihypothesis map-matching using particle filtering. 16th World Congress of ITS Systems and Services, Stockholm, Sweden
- Bouju A, Stockus A, Bertrand F, Boursier P (2002) Location-based spatial data management in navigation systems. *IEEE Symp Intell Veh* 1:172–177
- Czerniak RJ, Reilly JP (1998) NCHRP synthesis of highway practice 258: applications of GPS for surveying and other positioning needs in departments of transportation. Transportation Research Board, National Research Council, Washington, D.C
- Ding L, Chi L, Chen JB, Song C (2007) Improved neural network information fusion in integrated navigation system. In: *Proceedings of first IEEE international conference on mechatronics and automation*, Harbin, China, 5–8 Aug 2007
- Drane C, Rizos C (1998) Positioning systems in intelligent transportation systems. Artech House, London
- Feng S, Ochieng WY (2007) Integrity of navigation system for road transport. In: *Proceedings of the 14th ITS world congress*, Beijing, China, 9–13 Oct 2007
- Fuchs H, Kedem ZM, Naylor BF (1980) On visible surface generation by a priori tree structures. *Comput Graph* 14:124–133
- Goodwin C, Lau J (1993) Vehicle navigation and map quality. In: *Proceedings of the IEEE-IEE vehicle navigation & information systems conference*, Ottawa, pp 17–20
- Greenfeld JS (2002) Matching GPS observations to locations on a digital map. In: *Proceedings of the 81st annual meeting of the transportation research board*, Washington D.C., Jan 2002

- Jabbour M, Bonnifait P, Cherfaoui V (2008) Map-matching integrity using multihypothesis road-tracking. *J Intell Transport syst* 12(4):189–201
- Jo T, Haseyamai M, Kitajima H (1996) A map matching method with the innovation of the Kalman filtering. *IEICE Trans Fundam Electron Commun Comput Sci* E79-A:1853–1855
- Kaplan ED, Hegarty CJ (2006) *Understanding GPS – principles and applications*, 2nd edn. Artech House, Boston/London
- Kim S, Kim J (2001) Adaptive fuzzy-network based C-measure map matching algorithm for car navigation system. *IEEE Trans Ind Electron* 48(2):432–440
- Kim JS, Lee JH, Kang TH, Lee WY, Kim YG (1996) Node based map matching algorithm for car navigation system. In: *Proceeding of the 29th ISATA symposium*, vol 10. Florence, pp 121–126
- Kim W, Jee G, Lee J (2000) Efficient use of digital road map in various positioning for ITS. In: *Proceedings of IEEE symposium on position location and navigation*, San Deigo, CA
- Krakiwsky EJ, Harris CB, Wong RVC (1988) A Kalman filter for integrating dead reckoning, map matching and GPS positioning. In: *Proceedings of IEEE position location and navigation symposium*, pp 39–46
- Lee J, Won DH, Sung S, Kang TS, Lee YJ (2007) High assurance GPS integrity monitoring system using particle filtering approach. In: *Proceedings of 10th IEEE high assurance systems engineering symposium*, pp 437–438
- Li J, Fu M (2003) Research on route planning and map-matching in vehicle GPS/deadreckoning/electronic map integrated navigation system. *IEEE Proc Intell Transport Syst* 2:1639–1643
- Najjar ME, Bonnifait P (2005) A roadmap matching method for precise vehicle localization using belief theory and Kalman filtering. *Auton Robot* 19:173–191
- Nassreddine G, Abdallah F, Denreux T (2008) Map matching algorithm using belief function theory. In: *Proceedings of the 11th international conference on information fusion*
- NRC, National Research Council (2002) *Collecting, processing and integrating GPS data with GIS. NCHRP synthesis 301*. National Academy Press, Washington D.C
- Ochieng WY, Feng S, Moore T, Hill C, Hide C (2007) User level integrity monitoring and quality control for seamless positioning in all conditions and environments. In: *Proceedings of ION GNSS 20th international technical meeting of the satellite division*, Fort Worth, USA, 25–28 Sept 2007, pp 2573–2583
- Philipp AB, Zunker H (2005) Integrity hits the road, GPS world. Aster Pub. Corp., Eugene, pp 30–36, July 2005
- Phuyal BP (2002) Method and use of aggregated dead reckoning sensor and GPS data for map matching. In: *Proceedings of Institute of Navigation-GPS (ION-GPS) annual conference*, Portland, pp 430–437
- Pink O, Hummel B (2008) A statistical approach to map matching using road network geometry, topology and vehicular motion constraints. In: *Proceedings of the 11th international IEEE conference on intelligent transportation systems*, Beijing, China
- Pyo J, Shin D, Sung T (2001) Development of a map matching method using the multiple hypothesis technique. In: *IEEE proceedings on intelligent transportation systems*, pp 23–27
- Quddus MA (2006) High integrity map matching algorithms for advanced transport telematics applications. Ph.D. Thesis, Centre for transport studies, Imperial College London, UK
- Quddus MA, Ochieng WY, Lin Z, Noland RB (2003) A general map matching algorithm for transport telematics applications. *GPS Solut* 73:157–167
- Quddus MA, Ochieng WY, Noland RB (2006a) A high accuracy fuzzy logic-based map matching algorithm for road transport. *J Intell Transp Syst: Technol Plann Operat* 10:103–115
- Quddus MA, Ochieng WY, Noland RB (2006b) Integrity of map-matching algorithms. *Transp Res C: Emerg Technol* 144:283–302
- Quddus MA, Ochieng WY, Noland RB (2007) Current map matching algorithm for transport applications: a state-of-the art and future research direction. *Transp Res Part-C* 15:312–328
- Quddus MA, Noland RB, Ochieng WY (2008) The effects of navigation sensors and spatial road network data quality on the performance of map matching algorithms. *Geoinformatica* 13:85–108
- Scott CA, Drane CR (1994) Increased accuracy of motor vehicle position estimation by utilizing map data, vehicle dynamics and other

- information sources. In: Proceedings of the vehicle navigation and information systems conferences, pp 585–590
- Smaili C, Najjar ME, Charpillat F (2008) A road matching method for precise vehicle localization using hybrid bayesian network. *J Intell Transp Syst* 12(4):176–188
- Srinivasan D, Chue RL, Tan CW (2003) Development of an improved ERP system using GPS and AI technique. In: IEEE proceedings on intelligent transportation systems, pp 554–559
- Su H, Chen J, Xu J (2008) A adaptive map matching algorithm based on fuzzy-neural-network for vehicle navigation system. In: Proceedings of the 7th world congress on intelligent control and automation, Chongqing, China, 25–27 June 2008
- Syed S, Cannon ME (2004) Fuzzy logic-based map matching algorithm for vehicle navigation system in urban canyons. In: Proceedings of the institute of navigation (ION) national technical meeting, California, USA, 26–28 Jan 2004
- Taghipour S, Meybodi MR, Taghipour A (2008) An algorithm for map matching for car navigation system. In: Proceedings of 3rd IEEE international conference on information and communication technologies: From theory to application
- Tanaka J, Hirano K, Itoh T, Nobuta H, Tsunoda S (1990) Navigation system with map-matching method. In: Proceeding of the SAE international congress and exposition, pp 40–50
- Taylor G, Blewitt G, Steup D, Corbett S, Car A (2001) Road reduction filtering for GPS-GIS navigation. *Trans GIS* 5(3):193–207
- Torriti MT, Guesalaga A (2008) Scan-to-map matching using the Hausdorff distance for Robust mobile robot localization. In: Proceedings of 2008 IEEE international conference on robotics and automation, Pasadena, CA, USA
- Urmson C, Anhalt J, Bagnell D (2008) Autonomous driving in urban environments: boss and the urban challenge. *J Field Rob* 25(8):425–466
- Velaga NR, Quddus MA, Bristow AL (2009) Developing an enhanced weight based topological map-matching algorithm for intelligent transport systems. *Transp Res Part-C* 17:672–683
- Velaga NR, Quddus MA, Bristow AL (2010a) Detecting and correcting map-matching errors in location-based intelligent transport systems, 12th world conference on transport research, Lisbon, Portugal, 11–15 July 2010
- Velaga NR, Quddus MA, Bristow AL, Zheng Y (2010b) Map-aided integrity monitoring of a land vehicle navigation system. In: Proceedings of the 89th annual meeting of the transportation research board, Washington D.C., Jan 2010
- White CE, Bernstein D, Kornhauser AL (2000) Some map-matching algorithms for personal navigation assistants. *Transp Res C* 8:91–108
- Xu H, Liu H, Norville HS, Bao Y (2007) A virtual differential map-matching algorithm. In: Proceedings of the IEEE intelligent transportation systems conference seattle, WA, USA
- Yang D, Cai B, Yuan Y (2003) An improved map-matching algorithm used in vehicle navigation system. In: Proceedings of IEEE conference on intelligent transportation systems, pp 1246–1250
- Young MS, Stanton NA, Harris D (2007) Driving automation: learning from aviation about design philosophies. *Int J Veh Des* 45(3):323–338
- Yu M, Li Z, Chen Y, Chen W (2006) Improving integrity and reliability of map matching techniques. *J Global Positioning Syst* 5:40–46
- Zhang Y, Gao Y (2008) An improved map matching algorithm for intelligent transportation system. In: Proceedings of IEEE international conference on industrial technology, pp 1–5
- Zhao Y (1997) Vehicle location and navigation systems. Artech House, Inc., Norwood

15 Situational Awareness and Road Prediction for Trajectory Control Applications

Christian Lundquist · Thomas B. Schön · Fredrik Gustafsson
Department of Electrical Engineering, Division of Automatic Control, Linköping University, Linköping, SE, Sweden

1	<i>Introduction</i>	366
2	<i>Modeling the Environment with a Map</i>	367
3	<i>Feature-Based Map</i>	369
3.1	Radar and Laser	370
3.2	Cameras and Computer Vision	370
4	<i>Road Map</i>	373
4.1	Road Model	374
4.2	Mapping of the Road Lanes	377
4.3	Mapping of the Road Edges	380
5	<i>Occupancy Grid Map</i>	386
5.1	Background	386
5.2	OGM with Radar Measurements	388
5.3	Experiments and Results	388
6	<i>Intensity-Based Map</i>	389
7	<i>Conclusion</i>	394

Abstract: Situational awareness is of paramount importance in all advanced driver assistance systems. Situational awareness can be split into the tasks of tracking moving vehicles and mapping stationary objects in the immediate surroundings of the vehicle as it moves. This chapter focuses on the map estimation problem. The map is constructed from sensor measurements from radars, lasers and/or cameras, with support from on-board sensors for compensating for the ego-motion.

Four different types of maps are discussed:

- (i) Feature-based maps are represented by a set of salient features, such as tree trunks, corners of buildings, lampposts and traffic signs.
- (ii) Road maps make use of the fact that roads are highly structured, since they are built according to clearly specified road construction standards. This allows relatively simple and powerful models of the road to be employed.
- (iii) Location-based maps consist of a grid, where the value of each element describes the property of the specific coordinate.
- (iv) Finally, intensity-based maps can be considered as a continuous version of the location-based maps.

The aim is to provide a self-contained presentation of how these maps can be built from measurements. Real data from Swedish roads are used throughout the chapter to illustrate the methods.

1 Introduction

Most automotive original equipment manufacturers today offer longitudinal control systems, such as adaptive cruise control (ACC) or collision mitigation systems. Lateral control systems, such as lane keeping assistance (LKA), emergency lane assist (ELA) (Eidehall et al. 2007) and curve speed warning, are currently developed and released. These systems can roughly be split into safety applications, which aim to mitigate vehicular collisions such as rear end or blind spot detection; and comfort applications such as ACC and LKA, which aim at reducing the driver's work load. The overview article by Caveney (2010) describes the current development of trajectory control systems. The requirements on the position accuracy of the ego vehicle in relation to other vehicles, the road, and the surrounding environment increases with those control applications that are currently under development and expected to be introduced to the market.

The systems available or in development today are based on two basic tracking and decision principles: longitudinal systems use a radar, possibly supported by a camera, to track leading vehicles, and they decide on braking warnings or interventions. On the other hand, lateral systems use a camera to track the lane markers, and they decide on steering warnings or interventions. Future safety and comfort functions need more sophisticated situational awareness and decision functionality:

- A combination of lateral and longitudinal awareness will be needed, where all lanes are monitored, all of their vehicles are tracked, and the road-side conditions are modeled

to allow for emergency maneuvers. The result is a situational awareness map, which is the topic for this chapter.

- This will allow for more sophisticated decision functionality. First, the possible evasive driver maneuvers are computed, and only if the driver has no or very little time for evasive actions, the system will intervene. Second, more complex automatic evasive maneuvers can be planned using the situational awareness map, including consecutive lateral and braking actions.

It should be remarked that the accuracy of the navigation systems today and in the near future (see ❶ Chaps. 17 and ❷ 18) is not of much assistance for situational awareness. The reason is that satellite-based navigation gives an accuracy of 10–20 m, which is not sufficient for lateral awareness. Even in future systems, including reference base stations, enabling meter accuracy, the standard road maps will limit the performance since they are not of sufficient accuracy. Thus, two leaps in development are needed before positioning information and standard maps can be used to improve situational awareness maps. Another technical enabler is car to car communication (C2C), which may improve tracking of other vehicles and in the end change the transportation system as has already been done with the transponder systems for aircraft and commercial surface ships. Still, there will always be vehicles and obstacles without functioning communication systems. The need for accurate situation awareness and road prediction to be able to automatically position the car in a lane and derive drivable trajectories will evolve and remain important.

The different types of situation awareness maps used to represent the environment are introduced in ❶ Sect. 2. Details of these maps are presented in ❶ Sects. 3–6. The chapter is concluded in ❶ Sect. 7.

2 Modeling the Environment with a Map

The transportation system may be described and represented by a number of variables. These variables include state variables describing the position, orientation, velocity, and size of the vehicles. Here one can distinguish between the own vehicle, the so-called ego vehicle, and the other vehicles, referred to as the targets.

The state variable of the ego vehicle at time k is denoted $\mathbf{x}_{E,k}$. The trajectory of the ego vehicle is recorded in $\mathbf{x}_{E,1:k} = \{\mathbf{x}_{E,1}, \dots, \mathbf{x}_{E,k}\}$, and it is assumed to be a priori known in this work. This is a feasible assumption since the absolute trajectory in world coordinates and the relative position in the road network are separable problems.

The state variable of the targets at time k is denoted $\mathbf{x}_{T,k}$. The road and the environment may be modeled by a map, which is represented by a set of variables describing N_m landmarks in the environment according to

$$\mathbf{M}_k = \{\mathbf{m}_k^{(1)}, \mathbf{m}_k^{(2)}, \dots, \mathbf{m}_k^{(N_m)}\}. \quad (15.1)$$

According to Thrun et al. (2005), there exist primarily two types of indexing for probabilistic maps. In a *feature-based map*, each $\mathbf{m}^{(n)}$ specifies the properties and location of one object, whereas in a *location-based map*, the index n corresponds to a location and $\mathbf{m}^{(n)}$ is the property of that specific coordinate. The *occupancy grid map* is a classical location-based representation of a map, where each cell of the grid is assigned a binary occupancy value that specifies if the location n is occupied ($\mathbf{m}^{(n)} = 1$) or not ($\mathbf{m}^{(n)} = 0$); see e.g., Elfes (1987), Moravec (1988).

The ego vehicle perceives information about the other vehicles and the environment through its sensors. The sensors provide a set of noisy measurements

$$\mathbf{Z}_k = \left\{ \mathbf{z}_k^{(1)}, \mathbf{z}_k^{(2)}, \dots, \mathbf{z}_k^{(N_{z,k})} \right\} \quad (15.2)$$

at each discrete time instant $k = 1, \dots, K$. Common sensors used for automotive navigation and mapping measure either range and bearing angle, as, for example, radar and laser, or bearing and elevation angles, as for the case of a camera. A signal preprocessing is always included in automotive radar sensors, and the sensor provides a list of detected features, defined by the range r , range rate \dot{r} , and bearing ψ . The preprocessing, the waveform design, and the detection algorithms of the radar are well described by, e.g., Rohling and Möller (2008) and Rohling and Meinecke (2001). Laser sensors typically obtain one range measurement per beam, and there exist both sensors which emit several beams at different angles and those which have a rotating beam deflection system. They all have in common that the angles at which they measure range are quantized, thus providing a list of range and bearings of which only the ones which are less than the maximum range shall be considered. Another commonly used automotive sensor is the camera. The camera measurements are quantized and the data is represented in a pixel matrix as an image.

Note that the indexing of sensor data is analogous to the representation of maps. Each range and bearing measurement $\mathbf{z}^{(n)}$ from a radar or laser specifies the properties and location of one observation, i.e., it is a feature-based measurement. However, the indexing of camera measurement is location based since the index n corresponds to a pixel in the image and $\mathbf{z}^{(n)}$ is the property of that specific coordinate.

The aim of all stochastic mapping algorithms, independent of indexing, is to estimate the posterior density of the map

$$p(\mathbf{M}_k | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k}), \quad (15.3)$$

given all the measurements $\mathbf{Z}_{1:k}$ from time 1 to k and the trajectory of the ego vehicle $\mathbf{x}_{E,1:k}$. The conditioning on $\mathbf{x}_{E,1:k}$ is implicit in this chapter since it is assumed to be a priori known. To be able to estimate the map, a relation between the map and the measurements must first be found and modeled. This model h is referred to as the measurement equation and for one combination of a measurement $\mathbf{z}_k^{(i)}$ and a map variable $\mathbf{m}_k^{(j)}$ it may be written according to

$$\mathbf{z}_k^{(i)} = h(\mathbf{m}_k^{(j)}) + \mathbf{e}_k, \quad (15.4)$$

where \mathbf{e}_k is the measurement noise. The primary aim in this chapter is to create a momentary map of the environment currently surrounding the ego vehicle. Hence, it is just the present map data that is recorded in the vehicle. As soon as a part of the environment is sufficiently far from the ego vehicle, the corresponding map entries are deleted. Environmental models must be compact so that they can be transmitted to and used efficiently by other automotive systems, such as path planners. The maps must be adapted to the type of environment they aim to model. For this reason, four different map representations, which are relevant for modeling the environment surrounding a vehicle, are described in the following sections.

Feature-based map	The map is represented by a number of salient features in the scene. Feature representation and tracking as part of a map is described in Sect. 3
Road map	This is a special case of the feature-based map, where the map variables model the geometry of the road. Roads are highly structured; they are built according to road construction standards and contain primarily straight lines, curves, and clothoids. Maps of road lanes and edges are described in Sect. 4
Location-based map	One of the most well-established location-based map is the occupancy grid map, which is described in Sect. 5 . The map is defined over a continuous space, but discretized with a grid approximation
Intensity-based map	The intensity density may be interpreted as the probability that one object is located in an infinitesimal region of the state space. The intensity-based map is a continuous approximation of a location-based map and it is described in Sect. 6

The estimated maps can be used to increase the localization accuracy of the ego vehicle with respect to its local environment. Furthermore, the maps may be used to derive a trajectory, which enables a collision free path of the vehicle.

3 Feature-Based Map

Features corresponding to distinct objects in the physical world, such as tree trunks, corners of buildings, lampposts, and traffic signs are commonly denoted *landmarks*. The procedure of extracting features reduces the computational complexity of the system as the features are on a more compact format than the original measurement. The form of the measurement equations ([15.4](#)) depends on the type of sensor used, and the signals measured by the sensor. In this section, we will briefly describe the use of features and the corresponding measurement equations in both radar and laser sensors ([Sect. 3.1](#)) as well as cameras ([Sect. 3.2](#)).

The feature-based approach may together with existing road maps be used to supplement the GPS-based position of the vehicle. This approach is also commonly referred to as visual odometry; see, e.g., Nistér et al. (2006).

3.1 Radar and Laser

As mentioned in the introduction, radar and laser sensors measure at least range and bearing of the landmark relative to the vehicles local coordinate, i.e., the measurement vector is composed of

$$\mathbf{z}^{(i)} = [r^{(i)} \psi^{(i)}]^T. \quad (15.5)$$

Notice that, for the sake of simplicity, the subscripts k specifying the time stamps of the quantities is dropped throughout this chapter. The assumption will be made that the measurements of the features are independent, i.e., the noise in each individual measurement $\mathbf{z}^{(i)}$ is independent of the noise in the other measurements $\mathbf{z}^{(j)}$, for $i \neq j$. This assumption makes it possible to process one feature at a time in the algorithms. Assume that the ego vehicle pose is defined by

$$\mathbf{x}_E = [x_E \ y_E \ \psi_E]^T, \quad (15.6)$$

where x_E, y_E denote the horizontal position of the vehicle and ψ_E denotes the heading angle of the vehicle. Furthermore, let us assume that one feature j in the map is defined by its Cartesian coordinate,

$$\mathbf{m}^{(j)} = [x_m^{(j)} \ y_m^{(j)}]^T. \quad (15.7)$$

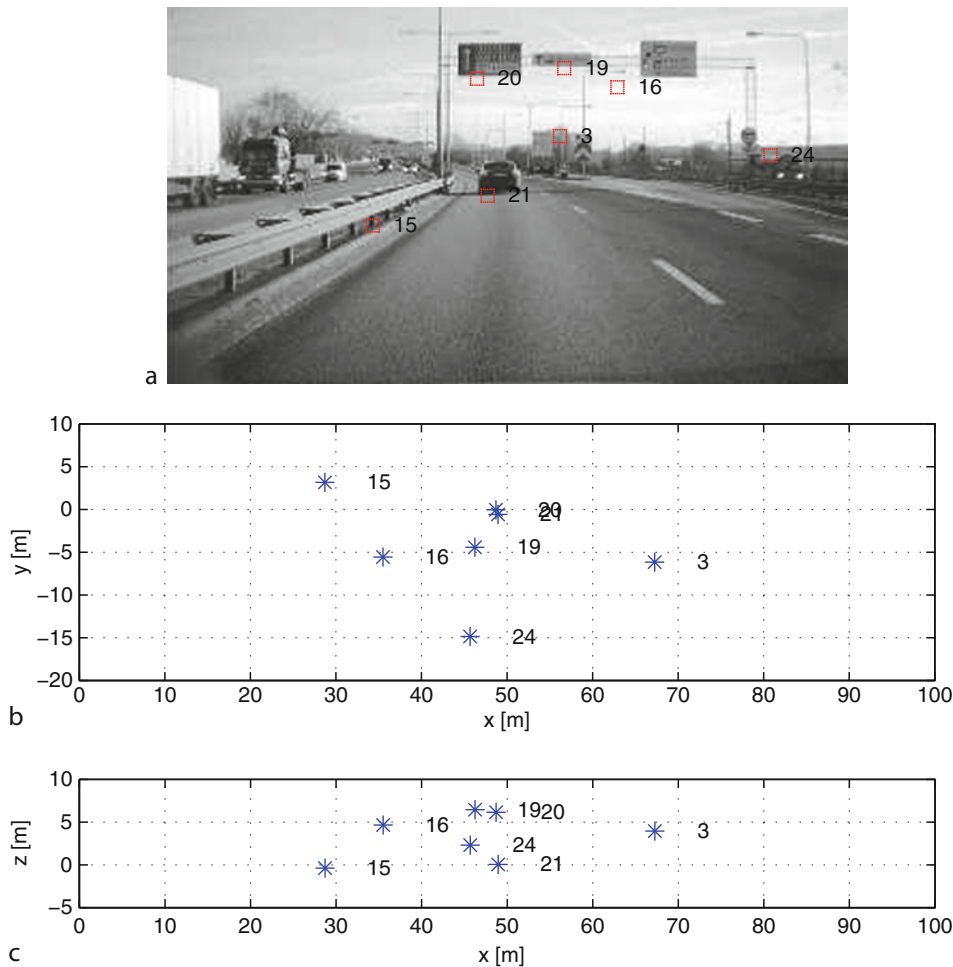
The measurement model ([15.4](#)) is then written as

$$\begin{bmatrix} r^{(i)} \\ \psi^{(i)} \end{bmatrix} = \begin{bmatrix} \sqrt{(x_m^{(j)} + x_E)^2 + (y_m^{(j)} + y_E)^2} \\ \arctan \frac{y_m^{(j)} - y_E}{x_m^{(j)} + x_E} - \psi_E \end{bmatrix} + \begin{bmatrix} e_r \\ e_\psi \end{bmatrix}, \quad (15.8)$$

where e_r and e_ψ are noise terms. The i th measurement feature corresponds to the j th map feature. The data association problem arises when the correspondence between the measurement feature and the map feature cannot be uniquely identified. A correspondence variable, which describes the relation between measurements and map features is introduced. This variable is also estimated at each time step k , and there exist a number of different algorithms to do this; see, e.g., Blackman and Popoli (1999). There are quite a few different possibilities on how to define features when radars and lasers are used.

3.2 Cameras and Computer Vision

Features form the bases in many computer vision algorithms, especially when it comes to building maps. There exists a myriad of feature detectors, which extract edges, corners or other distinct patterns. Some of the most well known are the Harris corner detector (Harris and Stephens 1988), SIFT (Lowe 2004), and MSER (Matas et al. 2004); see [Fig. 15.1](#) for an example, where the Harris corner detector is used. For a more complete



■ Fig. 15.1
Features detected using the Harris corner detector are shown in (a). (b) and (c) shows the estimated position of the landmarks in the x–y and x–z plane, respectively

account of various features used, see e.g., Szeliski (2010). Using features to build maps from camera images has been studied for a long time, and a good account of this is provided by Davison et al. (2007).

A key component in building maps using features is a good mathematical description of how the features detected in the image plane are related to the corresponding positions in world coordinates. The distance (commonly referred to as the depth) to a landmark cannot be determined from a single image, and this fact should be encoded by the mathematical parameterization of the landmark. The so-called inverse depth parameterization by Civera et al. (2008) provides an elegant uncertainty description of the fact that

the depth (i.e., distance) to the landmark is unknown. Here, the landmark (lm) state vector is given by

$$\mathbf{m}^{(j)} = \begin{bmatrix} \mathbf{c}^{(j)} \\ \psi^{(j)} \\ \phi^{(j)} \\ \rho^{(j)} \end{bmatrix} = \begin{bmatrix} \text{camera position first time lm was seen} \\ \text{azimuth angle of lm seen from } \mathbf{c}^{(j)} \\ \text{elevation angle of lm seen from } \mathbf{c}^{(j)} \\ \text{inverse distance (depth) from } \mathbf{c}^{(j)} \text{ to lm} \end{bmatrix} \quad (15.9a)$$

$$\mathbf{c}^{(j)} = \begin{bmatrix} x^{(j)} & y^{(j)} & z^{(j)} \end{bmatrix}^T, \quad (15.9b)$$

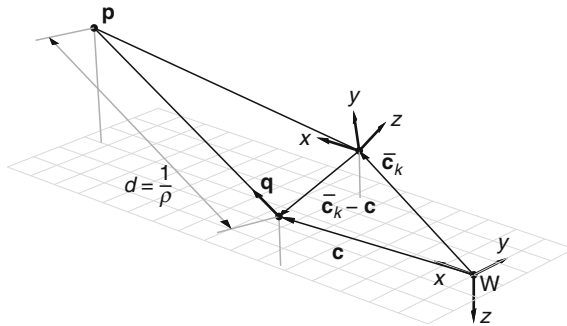
where $\mathbf{c}^{(j)}$ is the position of the camera expressed in world coordinates at the time when landmark j was first seen, $\psi^{(j)}$ is the azimuth angle of the landmark as seen from $\mathbf{c}^{(j)}$, relative to the world coordinate frame. The elevation angle of the landmark as seen from $\mathbf{c}^{(j)}$, relative to world coordinate frame directions is denoted $\phi^{(j)}$, and the inverse depth, which is the inverse of the distance, from $\mathbf{c}^{(j)}$ to the landmark is denoted $\rho^{(j)}$.

The landmark state $\mathbf{m}^{(j)}$ is a parametrization of the Cartesian position $\mathbf{p}^{(j)}$ of landmark j ; see [Fig. 15.2](#). The relationship between the position of the landmark and the inverse depth state representation is given by

$$\mathbf{p}^{(j)} = \mathbf{c}_k^{(j)} + \frac{1}{\rho^{(j)}} \underbrace{\begin{bmatrix} \cos \phi^{(j)} \cos \psi^{(j)} \\ \cos \phi^{(j)} \sin \psi^{(j)} \\ \sin \phi^{(j)} \end{bmatrix}}_{\mathbf{q}^{(j)}}. \quad (15.10)$$

The measurement model ([15.4](#)) for the landmarks is given by

$$h(\mathbf{m}^{(j)}) = P_n(\mathbf{p}^{C,(j)}) = \frac{1}{x_p^{C,(j)}} \begin{bmatrix} y_p^{C,(j)} \\ z_p^{C,(j)} \end{bmatrix}, \quad (15.11)$$



■ Fig. 15.2

The inverse depth parameterization used for the landmarks. The position of the landmark \mathbf{p} is parameterized using the position \mathbf{c} of the camera the first time the feature was seen, the direction $\mathbf{q}(\phi, \psi)$ and the inverse depth ρ . The position of the camera at time step k is denoted \mathbf{c}_k

where $P_n(\mathbf{p}^{C,(j)})$ is used to denote the normalized pinhole projection and $\begin{bmatrix} x_p^{C,(j)} & y_p^{C,(j)} & z_p^{C,(j)} \end{bmatrix}^T$ denotes the position of feature j at time k in the camera coordinate frame C . Note that before an image position can be used as a measurement together with the measurement equation (15.11), the image position is adjusted according to the camera specific parameters, such as focal length, pixel sizes, etc. The transformation between pixel coordinates $[u \ v]^T$ and normalized camera coordinates $[y \ z]^T$, which is the kind of coordinates landmark measurements $\mathbf{z}^{(i)}$ (see (15.11)) are given in, is

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} \frac{u-u^{ic}}{f_u} \\ \frac{v-v^{ic}}{f_v} \end{bmatrix}, \quad (15.12)$$

where $[u^{ic} \ v^{ic}]^T$ denotes the image center, and f_u and f_v are the focal lengths (given in pixels) in the lateral u direction and the vertical v direction, respectively. The transformation between world and camera coordinates is given by

$$\mathbf{p}^{C,(j)} = R^{CE} \left(R^{EW} \left(\begin{bmatrix} \mathbf{c}^{(j)} - \begin{bmatrix} x_E \\ y_E \end{bmatrix} \end{bmatrix} + \frac{1}{\rho^{(j)}} \mathbf{q}^{(j)} \right) - \mathbf{c}^E \right), \quad (15.13a)$$

$$R^{CE} = R(\psi^C, \phi^C, \gamma^C), \quad (15.13b)$$

$$R^{EW} = R(\psi_E, \phi_E, \gamma_E), \quad (15.13c)$$

where \mathbf{c}^E denotes the position of the camera expressed in the ego vehicle body coordinate frame. The rotation matrix $R(\alpha, \beta, \gamma)$ transforms coordinates from coordinate frame B to coordinate frame A , where the orientation of B relative to A is ψ (yaw), ϕ (pitch), and γ (roll). Furthermore, ψ^C , ϕ^C , and γ^C are the constant yaw, pitch, and roll angles of the camera, relative to the vehicle body coordinate frame.

The landmarks are estimated recursively using, e.g., a Kalman filter. An example of estimated landmarks is shown in Fig. 15.1. The estimated position $\mathbf{p}^{(j)}$ for seven landmarks is shown in the image plane as well as in the world x – y and x – z plane, where the ego vehicle is in the origin.

4 Road Map

A road map describe the shape of the road. The roads are mainly modeled using polynomial functions which describe the lane and the road edges. The advantages of road models are that they require sparse memory and are still very accurate, since they do not suffer from discretization problems. General road models are presented in Sect. 4.1. Lane estimation using camera measurements is described in Sect. 4.2 and finally road edge estimation based on feature measurements is described in Sect. 4.3.

4.1 Road Model

The road, as a construction created by humans, possesses no dynamics; it is a static time invariant object in the world coordinate frame. The building of roads is subject to road construction standards; hence, the modeling of roads is geared to these specifications. However, if the road is described in the ego vehicle's coordinate frame and the vehicle is moving along the road, it is possible and indeed useful to describe the characteristics of the road using time-varying state variables.

A road consists of straight and curved segments with constant radius and of varying length. The sections are connected through transition curves so that the driver can use smooth and constant steering wheel movements instead of stepwise changes when passing through road segments. More specifically, this means that a transition curve is formed as a clothoid, whose curvature c changes linearly with its curve length x_c according to

$$c(x_c) = c_0 + c_1 \cdot x_c. \quad (15.14)$$

Note that the curvature c is the inverse of the radius. Now, suppose x_c is fixed to the ego vehicle, i.e., $x_c = 0$ at the position of the ego vehicle. When driving along the road and passing through different road segments, c_0 and c_1 will not be constant, but rather time-varying state variables

$$\mathbf{m} = \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \text{curvature at the ego vehicle} \\ \text{curvature derivative} \end{bmatrix}. \quad (15.15)$$

In [Sect. 3](#), the map features were expressed in a fixed world coordinate frame. However, note that in this section, the road map is expressed as seen from the moving ego vehicle. Using [\(15.14\)](#), a change in curvature at the position of the vehicle is given by

$$\left. \frac{dc(x_c)}{dt} \right|_{x_c=0} = \dot{c}_0 = \frac{dc_0}{dx_c} \cdot \frac{dx_c}{dt} = c_1 \cdot v, \quad (15.16)$$

where v is the ego vehicle's velocity. Furthermore, the process model is given by

$$\begin{bmatrix} \dot{c}_0 \\ \dot{c}_1 \end{bmatrix} = \begin{bmatrix} 0 & v_x \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} + \begin{bmatrix} 0 \\ w_{c_1} \end{bmatrix}. \quad (15.17)$$

This model is referred to as the *simple clothoid model*, and it is driven by the process noise w_{c_1} . Note that the road is modeled in a road-aligned coordinate frame, with the components (x_c, y_c) , and the origin at the position of the ego vehicle. There are several advantages of using road-aligned coordinate frames, especially when it comes to the process models of the other vehicles on the same road, where these models are greatly simplified in road-aligned coordinates. However, the flexibility of the process model is reduced and basic dynamic relations such as Newton's and Euler's laws cannot be directly applied. The road model [\(15.14\)](#) is transformed into Cartesian coordinates (x, y) using

$$x(x_c) = \int_0^{x_c} \cos(\chi(x)) dx \approx x_c, \quad (15.18a)$$

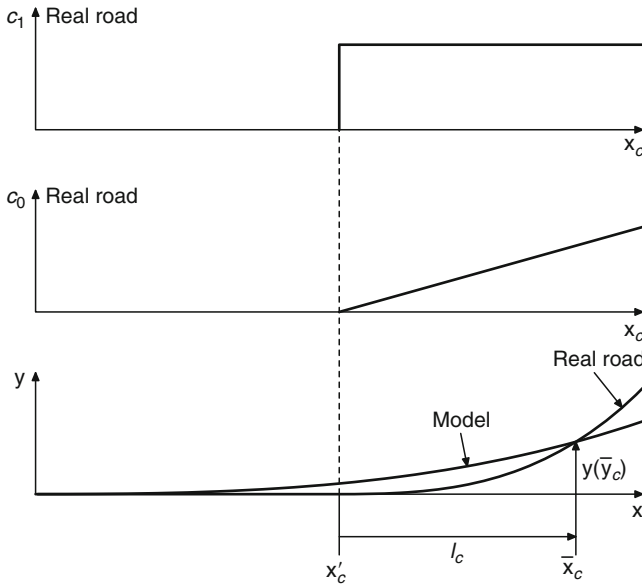
$$y(x_c) = \int_0^{x_c} \sin(\chi(x)) dx \approx \frac{c_0}{2} x_c^2 + \frac{c_1}{6} x_c^3, \quad (15.18b)$$

where the heading angle χ is defined as

$$\chi(x) = \int_0^x c(\lambda) d\lambda = c_0 x + \frac{c_1}{2} x^2. \quad (15.18c)$$

The origin of the two frames is fixed to the ego vehicle; hence, integration constants (x_0, y_0) are omitted.

A problem appears when two or more clothoid segments, with different parameters c_0 and c_1 , are observed in the same camera view. The parameter c_0 will change continuously during driving, whereas c_1 will be constant in each segment and change stepwise at the segment transition. This leads to a Dirac impulse in \dot{c}_1 at the transition. The problem can be solved by assuming a high process noise w_{c1} in (15.17), but this leads to less precise estimates of the state variables when no segment transitions occur in the camera view. To solve this problem, an averaging curvature model was proposed by Dickmanns (1988), which is perhaps best described with an example. Assume that the ego vehicle is driving on a straight road (i.e., $c_0 = c_1 = 0$) and that the look ahead distance of the camera is \bar{x}_c . A new segment begins at the position $x'_c < \bar{x}_c$, which means that there is a step in c_1 , and c_0 is ramped up; see Fig. 15.3. The penetration into the next segment is $l_c = \bar{x}_c - x'_c$. The idea of this model, referred to as *averaging* or *spread-out dynamic curvature model*, with



■ Fig. 15.3

A straight and a curved road segment are modeled with the averaging road model. The two upper plots show the parameters c_1 and c_0 of the real road, the bottom plot shows the real and the modeled roads in a Cartesian coordinate frame

the new state variables c_{0m} and c_{1m} , is that it generates the true lateral offset $y(\bar{x}_c)$ at the look ahead distance \bar{x}_c , i.e.,

$$y_{\text{real}}(\bar{x}_c) = y_{\text{model}}(\bar{x}_c), \quad (15.19)$$

but it is continuously spread out in the range $(0, \bar{x}_c)$. The lateral offset of the real road as a function of the penetration l_c , for $0 \leq l_c \leq \bar{x}_c$, is

$$y_{\text{real}}(l_c) = \frac{c_1}{6} l_c^3, \quad (15.20)$$

since the first segment is straight. The lateral offset of the averaging model as a function of the penetration l_c is

$$y_{\text{model}}(l_c) = \frac{c_{0m}(l_c)}{2} \bar{x}_c^2 + \frac{c_{1m}(l_c)}{6} \bar{x}_c^3, \quad (15.21)$$

at the look ahead distance \bar{x}_c . The equation

$$c_1 \frac{l_c^3}{\bar{x}_c^2} = 3c_{0m}(l_c) + c_{1m}(l_c) \bar{x}_c, \quad (15.22)$$

is obtained by inserting (15.20) and (15.21) into (15.19). By differentiating (15.22) with respect to l_c and using the relations $\frac{dc_1}{dl_c} = 0$, $\frac{dc_{0m}(l_c)}{dl_c} = c_{1m}(l_c)$ and $\frac{d(\cdot)}{dl_c} = \frac{d(\cdot)}{dt} \cdot \frac{dt}{dl_c}$, the following equation is obtained

$$\dot{c}_{1m} = 3 \frac{v}{\bar{x}_c} \left(c_1 (l_c / \bar{x}_c)^2 - c_{1m} \right), \quad (15.23)$$

for $l_c < \bar{x}_c$. Since $(l_c / \bar{x}_c)^2$ is unknown, it is usually set to 1 (Dickmanns 2007), which finally yields

$$\dot{c}_{1m} = 3 \frac{v}{\bar{x}_c} (c_1 - c_{1m}). \quad (15.24)$$

The state variable vector of the averaging model is defined as

$$\mathbf{m} = \begin{bmatrix} c_{0m} \\ c_{1m} \\ c_1 \end{bmatrix} = \begin{bmatrix} \text{curvature at the ego vehicle} \\ \text{averaged curvature derivative} \\ c \text{ derivative of the foremost segment} \end{bmatrix}, \quad (15.25)$$

and the process model is given by augmenting the simple clothoid model (15.17) with (15.24) according to

$$\begin{bmatrix} \dot{c}_{0m} \\ \dot{c}_{1m} \\ \dot{c}_1 \end{bmatrix} = \begin{bmatrix} 0 & v & 0 \\ 0 & -3 \frac{v}{\bar{x}_c} & -3 \frac{v}{\bar{x}_c} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c_{0m} \\ c_{1m} \\ c_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w_{c1} \end{bmatrix}. \quad (15.26)$$

The model is driven by the process noise w_{c1} , which also influences the other states. The averaging model is well described in the recent textbook Dickmanns (2007), and some early results using the model are presented by, e.g., Dickmanns and Mysliwetz (1992).

4.2 Mapping of the Road Lanes

The problem of mapping road lanes or lane estimation, as it is often called, is a *curve estimation* problem. The task is to obtain the best possible estimate of the curve describing the lane by exploiting the measurements provided by the onboard sensors. The most important sensor type here is exteroceptive sensors, such as, for example, cameras and lasers. Currently, the camera is the most commonly used sensor for the lane estimation problem, and in this section we will show how camera images can be used to obtain lane estimates.

The lane estimation problem is by no means a new problem; it has been studied for more than 25 years (see, e.g., Waxman et al. (1987), Dickmanns and Mysliwetz (1992) for some early work and Wang et al. (2008) for more recent contributions). A complete overview of what has been done on this problem is available in the survey paper McCall and Trivedi (2006). In this section, the problem is broken down into its constituents, and one way of solving the lane estimation problem when a camera is used as a sensor is shown.

The lane estimation problem can be separated into two subproblems, commonly referred to as the *lane-detection* problem and the *lane-tracking* problem. As the name reveals, the lane-detection problem deals with detecting lanes in an image. The lane-tracking problem then makes use of the detected lanes together with information about the temporal and the spatial dependencies over time in order to compute lane estimates. These dependencies are mathematically described using a road model. Traditionally, lane tracking is done using an extended Kalman filter; see, e.g., Dickmanns and Mysliwetz (1992), Guiducci (1999). There are also interesting approaches based on the particle filter by Gordon et al. (1993) available; see, e.g., Zhou et al. (2006), Wang et al. (2008), Kim (2008).

The lane is here modeled in the image plane (cf. [Sect. 3.2](#)) as a linear function close to the ego vehicle and as a quadratic function far away, i.e.,

$$l_{\theta}(v) = \begin{cases} a + b(v - v_s) & v > v_s \\ a + b(v - v_s) + c(v - v_s)^2 & v \leq v_s, \end{cases} \quad (15.27a)$$

where v_s denotes the (known) vertical separation in pixels between the linear and the quadratic model (illustrated by the horizontal line in [Fig. 15.4](#)) and subindex θ is used to denote the dependence on the parameters

$$\theta = [a \ b \ c]^T, \quad (15.27b)$$

which are to be estimates. These parameters all have geometrical interpretations in terms of offset (a), local orientation (b), and curvature (c) of the lane in the image plane. The lane estimates (here, the estimates of the parameters θ in [\(15.27a\)](#) and [b\)](#)) carry important information about the states in the road model introduced in [Sect. 4.1](#), which are expressed in world coordinates (in contrast to pixel coordinates u, v). These road model states are typically what we are interested in, and we will return to this important connection at the end of this section.



■ Fig. 15.4

Lane estimation results (in *gray*) overlaid onto the camera image. From this figure, the lane model (● 15.27) is clearly illustrated; the model is linear for $v > v_s$ and quadratic for $v \leq v_s$. The method assumes that the road surface is flat and when this assumption is true, the results are good; see (a). However, when this assumption does not hold, the estimates are not that good on a longer horizon; see (b)

Given the fact that the problem of lane detection has been studied for more than 25 years, there are many ideas on how to solve this problem available. Rather than trying to give a complete account of all the different methods available, we will here be very specific and explain one way in which lanes can be detected and show some results on real sequences. The solution presented here is very much along the lines of Jung and Kelber (2005), Lee (2002).

The initial lane detection is performed using a linear lane model, which is found from the image using a combination of the edge distribution function (EDF) (Lee 2002)

and the Hough transform (Hough 1962, Duda and Hart 1972). The EDF is defined as the gradient orientation

$$\varphi(u, v) = \arctan \left(\frac{D_u}{D_v} \right), \quad (15.28)$$

where D_u and D_v are approximations of the gradient function

$$\nabla I(u, v) = \left[\frac{\partial I}{\partial u} \frac{\partial I}{\partial v} \right]^T \approx [D_u \ D_v]^T \quad (15.29)$$

for the gray scale image $I(u, v)$. The two largest peaks (α^l , α^r) of $\varphi(u, v)$ provide the most probable orientations of the lanes in the image. This is used to form an edge-image $g(u, v)$ as

$$g(u, v) = \begin{cases} |\nabla I(u, v)| \approx |D_u| + |D_v|, & |\varphi(u, v) - \alpha^l| < T_\alpha \text{ or } |\varphi(u, v) - \alpha^r| < T_\alpha \\ 0, & \text{Otherwise} \end{cases} \quad (15.30)$$

where T_α is a threshold, here $T_\alpha = 2^\circ$. Applying the Hough transform to the edge-image $g(u, v)$ provides two initial linear models $\mathbf{l}(v) = a + bv$, one for the left lane markings and one for the right lane markings. These models are used to form a region which will serve as the search region in the subsequent image. This region, which we refer to as the lane boundary region of interest (LBROI) is simply defined by extending the linear model w pixels to the right and w pixels to the left (here $w = 10$).

Given that an initial LBROI is found, the task is now to make use of this information in order to compute estimates of the parameters in the lane model (► 15.27a),

$$\theta = \left[(\theta^l)^T (\theta^r)^T \right]^T, \quad \text{where } \theta^l = [a^l \ b^l \ c^l]^T, \quad \theta^r = [a^r \ b^r \ c^r]^T, \quad (15.31)$$

where superscript l and r have been used to indicate the left lane marking and the right lane marking, respectively. Estimates of these parameters θ are obtained by solving a constrained weighted least squares problem for each image. The cost function is given by

$$V(\theta) = \sum_{i=1}^N \left(M_i^l (u_i^l - \mathbf{l}_{\theta^l}(v_i^l))^2 + (M_i^r (u_i^r - \mathbf{l}_{\theta^r}(v_i^r))^2 \right), \quad (15.32)$$

where N denotes the number of relevant pixels in the lateral u direction, \mathbf{l} denotes the lane model given in (► 15.27a), and M_i denotes the magnitude in the thresholded edge-image $g_2(u, v)$, $M_i = g_2(u_i, v_i)$, defined as

$$g_2(u, v) = \begin{cases} g(u, v), & g(u, v) \geq \frac{1}{2} M_{\text{mean}} \\ 0, & \text{otherwise} \end{cases} \quad (15.33)$$

where M_{mean} denotes the mean magnitude of the $g(u, v)$. Constraints are introduced in order to account for the fact that the right lane and the left lane are related to each other. This is modelled according to the following linear (in θ) inequality constraint

$$a^r - a^l + (b^r - b^l)(v_1 - v_m) + (c^r - c^l)(v_1 - v_m)^2 \leq \delta, \quad (15.34)$$

which states that the left and the right lanes cannot be more than δ pixels apart furthest away (at v_1) from the host vehicle. In other words, (15.34) encodes the fact that the left and the right lanes must have similar geometry in the sense that the quadratic parts in (15.27) are strongly related.

From (15.32–15.34), it is now clear that lane estimation boils down to a curve estimation problem, which here is quadratic in the unknown parameters θ . More specifically, inserting the lane model (15.27a) into the cost function (15.32) and writing the problem on matrix form results in a constrained weighted least squares problem on the form

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{2} \theta^T H \theta + f^T \theta \\ \text{s.t.} \quad & L \theta \leq \delta. \end{aligned} \quad (15.35)$$

This is a quadratic program (QP) implying that a global minimum will be found, and there are very efficient solvers available for these type of problems. Here, we have made use of a dual active set method according to Gill et al. (1991). (The QP code was provided by Dr. Adrian Wills at the University of Newcastle, Australia; see <http://sigpromu.org/quadprog>. This code implements the method described by Goldfarb and Idnani [1983] and Powell [1985].) An illustration of the lane estimation results is provided in Fig. 15.4. The estimate of θ is then used to form the LBROI for the new image, simply as region defined by $\mathbf{l}_\theta(v) \pm w$ for each lane.

The lane estimates that are now obtained as the solution to (15.35) can be expressed in world coordinates, seen from the ego vehicle, using geometrical transformation along the lines of what has already been described in Sect. 3. These transformations are discussed in detail in Guiducci (2000). Once the lane estimates are available in the world coordinates, they can be used as camera measurements in a *sensor fusion* framework to make a very important contribution to the estimate of map variables \mathbf{m} (i.e., (15.15) or (15.24)) in the road model (perhaps most importantly the curvature c_0 and the curvature derivative c_1) as it is derived in Sect. 4.1.

4.3 Mapping of the Road Edges

Feature-based measurements of landmarks along the road may be used to map the road edges. This section describes a method to track line shaped objects, such as guardrails using point measurements from radar, laser, or extracted camera features. Tracking point objects was covered in Sect. 3 and is not repeated here. The line shaped and curved guardrails are described using the polynomial road model (15.18a and b) and tracked as extended targets in a Cartesian frame. However, to allow a more general treatment of

the problem in this section, the extended targets are modeled using n th order polynomials given as

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n, \quad (15.36)$$

in the range $[x_{\text{start}}, x_{\text{end}}]$, where $\mathbf{m}_a \triangleq [a_0 \ a_1 \ \dots \ a_n]^T$ are the polynomial coefficients and $[x \ y]^T$ are planar Cartesian coordinates. Note that the coordinate y is a function of x and that the direction of the coordinate frame is chosen depending on the application in mind. The state vector of a map object j is defined as

$$\mathbf{m}^{(j)} \triangleq \left[(\mathbf{m}_a^j)^T x_{\text{start}}^j x_{\text{end}}^j \right]^T. \quad (15.37)$$

The map \mathbf{M} is modeled by a set of such polynomial shapes according to (15.1).

Suppose the 2-dimensional noisy feature-based sensor measurements are given in batches of Cartesian x and y coordinates as follows

$$\left\{ \mathbf{z}_k^{(i)} \triangleq \left[x^{(i)} y^{(i)} \right]_k^T \right\}_{i=1}^{N_{z,k}}, \quad (15.38)$$

for discrete time instants $k = 1, \dots, K$. In many cases in reality (e.g., radar, laser, and stereo vision cf. (15.5)) and in the practical application considered in this section, the sensor provides range r and azimuth angle ψ given as,

$$\left\{ \bar{\mathbf{z}}_k^{(i)} \triangleq \left[r^{(i)} \psi^{(i)} \right]_k^T \right\}_{i=1}^{N_{z,k}}. \quad (15.39)$$

In such a case, some suitable standard polar to Cartesian conversion algorithm is used to convert these measurements into the form (15.38).

The state model considered in this section is described, in general, by the state-space equations

$$\mathbf{m}_{k+1} = f(\mathbf{m}_k, \mathbf{u}_k) + \mathbf{w}_k, \quad (15.40a)$$

$$\mathbf{y}_k = h(\mathbf{m}_k, \mathbf{u}_k) + \mathbf{e}_k, \quad (15.40b)$$

where \mathbf{m} , \mathbf{u} , and \mathbf{y} denote the state, the input signal, and the output signal, while $\mathbf{w} \sim N(0, Q)$ and $\mathbf{e} \sim N(0, R)$ are the process and measurement noise, respectively. The use of an input signal \mathbf{u} is important in this framework. For the sake of simplicity, the tracked objects are assumed stationary, resulting in very simple motion models (15.40a).

A polynomial is generally difficult to handle in a filter since the noisy measurements are distributed arbitrarily along the polynomial. In this respect, the measurement models considered contain parts of the actual measurement vector as parameters. The methodology takes into account the errors caused by using the actual noisy measurements as model parameters. This scheme is an example of the so-called errors-in-variables (EIV) framework; see, e.g., Söderström (2007), Diversi et al. (2005), and Björck (1996).

The general convention in modeling is to make the definitions

$$\mathbf{y} \triangleq \mathbf{z}, \quad \mathbf{u} \triangleq \emptyset, \quad (15.41)$$

where \emptyset denotes the empty set meaning that there is no input. Notice that the subscripts k , specifying the time stamps of the quantities, is omitted for the sake of simplicity. In this setting, it is extremely difficult, if not impossible, to find a measurement model connecting the outputs \mathbf{y} to the states \mathbf{m}_a in the form of (15.40b). Therefore, other selections for \mathbf{y} and \mathbf{u} , need to be used. Here, the selection

$$\mathbf{y} \triangleq \mathbf{y}, \quad \mathbf{u} \triangleq \mathbf{x}. \quad (15.42)$$

is made. Although being quite a simple selection, this choice results in a rather convenient linear measurement model in the state partition \mathbf{m}_a ,

$$\mathbf{y} = H_a(\mathbf{u})\mathbf{m}_a + \mathbf{e}, \quad (15.43)$$

where $H_a(\mathbf{u}) = [1 \times \mathbf{x}^2 \dots \mathbf{x}^n]^T$. It is the selection in (15.42) rather than (15.41) that allows to use the standard methods in target tracking with clever modifications. Such a selection as (15.42) is also in accordance with the EIV representations, where measurement noise is present in both the outputs and inputs, i.e., the observation \mathbf{z} can be partitioned according to

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix}. \quad (15.44)$$

The measurement vector given in (15.38) is expressed in terms of a noise free variable \mathbf{z}_0 which is corrupted by additive measurement noise $\tilde{\mathbf{z}}$ according to

$$\mathbf{z} = \mathbf{z}_0 + \tilde{\mathbf{z}}, \quad \tilde{\mathbf{z}} \sim N(0, \Sigma_c), \quad (15.45)$$

where the covariance Σ_c can be decomposed as

$$\Sigma_c = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_y \end{bmatrix}. \quad (15.46)$$

Note that, in the case the sensor provides measurements only in polar coordinates (15.39), one has to convert both the measurement \mathbf{z} and the measurement noise covariance

$$\Sigma_p = \text{diag}(\sigma_d^2, \sigma_\delta^2) \quad (15.47)$$

into Cartesian coordinates. This is a rather standard procedure. Note that, in such a case, the resulting Cartesian measurement covariance Σ_c is, in general, not necessarily diagonal, and, hence, Σ_{xy} of (15.46) might be nonzero.

Since the model (15.43) is linear, the Kalman filter measurement update formulas can be used to incorporate the information in \mathbf{z} into the extended source state \mathbf{m}_a . An important question in this regard is what would be the measurement covariance of the measurement noise term \mathbf{e} in (15.43).

Neglecting the errors in the model parameters $H_a(\mathbf{u})$ can cause overconfidence in the estimates of recursive filters and can actually make data association difficult in tracking applications (by causing too small gates). A simple methodology is used to take the uncertainties in $H_a(\mathbf{u})$ into account in line with the EIV framework. Assuming that the elements of the noise free quantity \mathbf{z}_0 satisfy the polynomial equation exactly according to

$$\mathbf{y} - \tilde{\mathbf{y}} = H_a(\mathbf{u} - \tilde{\mathbf{u}})\mathbf{m}_a, \quad (15.48a)$$

$$\mathbf{y} - \tilde{\mathbf{y}} = [1 \ x - \tilde{x} \ (x - \tilde{x})^2 \cdots (x - \tilde{x})^n] \mathbf{m}_a, \quad (15.48b)$$

which can be approximated using a first order Taylor expansion resulting in

$$\mathbf{y} \approx H_a(\mathbf{u})\mathbf{m}_a - \tilde{H}_a(\mathbf{u})\tilde{x}\mathbf{m}_a + \tilde{\mathbf{y}} \quad (15.49a)$$

$$= H_a(\mathbf{u})\mathbf{m}_a + \tilde{h}_a(\mathbf{m}_a, \mathbf{u}) \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}, \quad (15.49b)$$

with

$$H_a(\mathbf{u}) = [1 \ x \ x^2 \cdots x^n], \quad (15.49c)$$

$$\tilde{H}_a(\mathbf{u}) = [0 \ 1 \ 2x \cdots nx^{n-1}], \quad (15.49d)$$

$$\tilde{h}_a(\mathbf{m}_a, \mathbf{u}) = [-a_1 - 2a_2x \cdots -na_nx^{n-1} \ 1]. \quad (15.49e)$$

Hence, the noise term \mathbf{e} of (15.43) is given by

$$\mathbf{e} = \tilde{\mathbf{y}} - \tilde{H}_a\tilde{x}\mathbf{m}_a = \tilde{h}(\mathbf{m}_a, \mathbf{u}) \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \quad (15.50)$$

and its covariance is given by

$$\begin{aligned} \Sigma_a &= E(\mathbf{e}\mathbf{e}^T) = \Sigma_y + \mathbf{m}_a\tilde{H}_a\Sigma_x\tilde{H}_a^T\mathbf{m}_a^T - 2\tilde{H}_a\Sigma_{xy} \\ &= \tilde{h}(\mathbf{m}_a, \mathbf{u})\Sigma_c\tilde{h}^T(\mathbf{m}_a, \mathbf{u}). \end{aligned} \quad (15.51)$$

Note that the EIV covariance Σ_a depends on the state variable \mathbf{m}_a , which is substituted by its last estimate in recursive estimation.

Up to this point, only the relation of the observation \mathbf{z} to the state component \mathbf{m}_a has been considered. It remains to discuss the relation between the observation and the start x_{start} and the end points x_{end} of the polynomial. The measurement information must only be used to update these components of the state if the new observations of the extended source lie outside the range of the polynomial. The following (measurement dependent) measurement matrix can be defined for this purpose:

$$H_{se} = \begin{cases} [1 \ 0] & \text{if } x \leq x_{\text{start},k|k-1} \\ [0 \ 1] & \text{if } x \geq x_{\text{end},k|k-1} \\ [0 \ 0] & \text{otherwise.} \end{cases} \quad (15.52)$$

The complete measurement model of an extended object can now be summarized by

$$\mathbf{z} = H\mathbf{m} + \mathbf{e}, \quad \mathbf{e} \sim N(0, R(\mathbf{m})), \quad (15.53a)$$

with

$$H = \begin{bmatrix} 0^{1 \times n} & H_{se} \\ H_0 & 0^{1 \times 2} \end{bmatrix}, \quad (15.53b)$$

$$R(\mathbf{m}) = \text{blkdiag}(\Sigma_x, \Sigma_a(\mathbf{m})). \quad (15.53c)$$

Put in words, if the x -component of a new measurement is closer to the sensor than the start point of the line x_{start} , it is considered in the measurement equation (► 15.52) and can be used to update this state variable. Analogously, if a new measurement is more distant than the end point of the line x_{end} , it is considered in (► 15.52). Further, if a measurement is in between the start and end point of the line, the measurement model is zero in (► 15.52) and there is no relation between this measurement and the state variables x_{start} or x_{end} .

Any model as, e.g., the standard constant velocity or the coordinated turn model may be used for the targets. For simplicity, it is assumed that the targets are stationary in this contribution, thus the process model on the form (► 15.40a) is linear and may be written

$$\mathbf{m}_{k+1} = F\mathbf{m}_k + \mathbf{w}_k. \quad (15.54)$$

To increase the flexibility of the extended object an assumption about the dynamic behavior of its size is made. The size of the extended object is modeled to shrink with a factor $0.9 < \lambda < 1$ according to

$$x_{\text{start},k+1} = x_{\text{start},k} + \lambda(x_{\text{end},k} - x_{\text{start},k}), \quad (15.55a)$$

$$x_{\text{end},k+1} = x_{\text{end},k} - \lambda(x_{\text{end},k} - x_{\text{start},k}), \quad (15.55b)$$

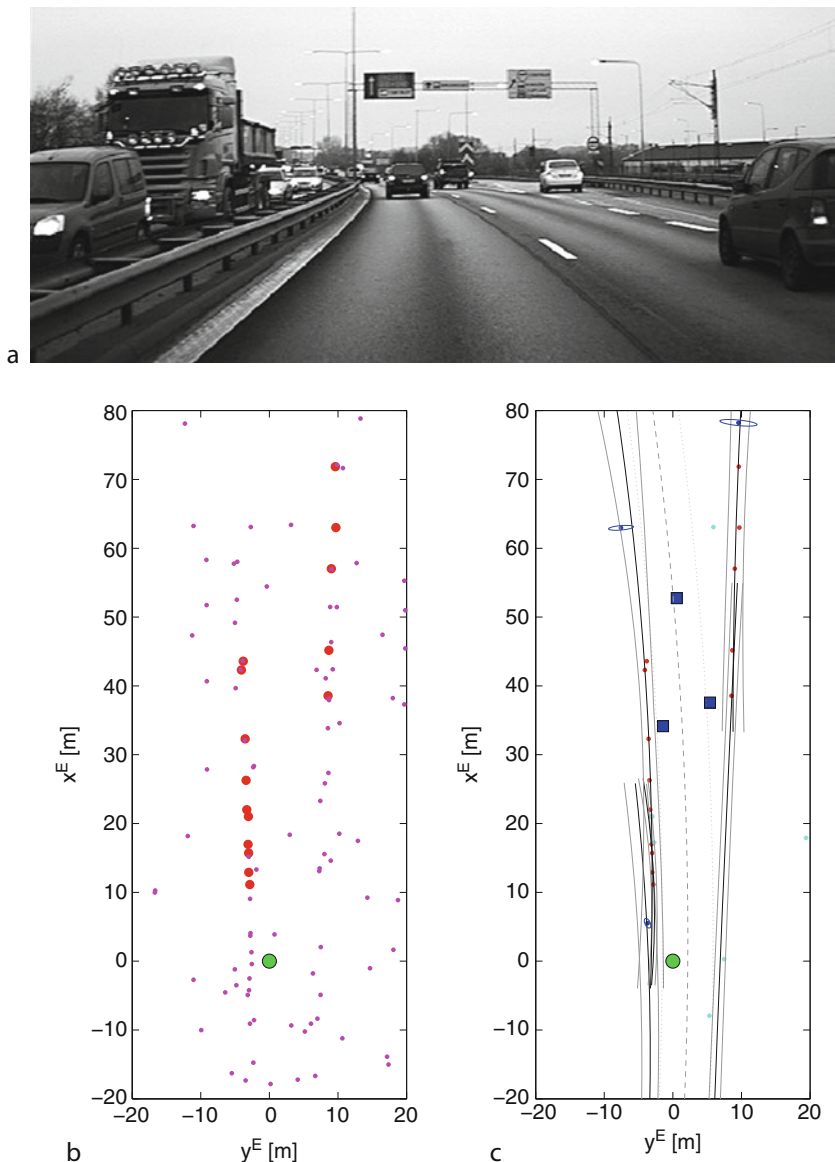
leading to the following process model for the polynomial

$$F = \begin{bmatrix} I^{n \times n} & 0^{n \times 1} \\ 0^{1 \times n} & 1 - \lambda & \lambda \\ & \lambda & 1 - \lambda \end{bmatrix}. \quad (15.56)$$

This shrinking behavior for the polynomials allows for automatic adjustment of the start and end points of the polynomials according to the incoming measurements.

The association of measurements to state estimates is treated in Lundquist et al. (2011b), where a generalized nearest neighbor method is applied.

The section is concluded with some results based on the information given by an ordinary automotive ACC radar, for the traffic situation shown in ► Fig. 15.5a. The ego vehicle, indicated by a circle, is situated at the (0,0)-position in ► Fig. 15.5b, and the dots are the radar reflections, or stationary observations, at one time sample. The smaller dots are former radar reflections, obtained at earlier time samples. ► Figure 15.5c



■ Fig. 15.5

A traffic situation is shown in (a). (b) shows the radar measurements, and (c), the resulting tracked points and lines. The *circle* in the origin is the ego vehicle, the *square* is the tracked vehicle in front, and the dashed *gray* lines illustrate the tracked road curvature

shows the estimated points and lines for the same scenario using the KF EIV method presented in this contribution. The mean values of the states are indicated by solid black lines or blue points. Furthermore, the state variance, by means of the 90% confidence interval, is illustrated by gray lines or cyan-colored ellipses, respectively. The estimate of

the lane markings (► 15.18a and b), illustrated by the gray dashed lines and derived according to the method presented in Lundquist and Schön (2011), is shown here as a comparison. Tracked vehicle in front of the ego vehicle are illustrated by blue squares.

5 Occupancy Grid Map

An occupancy grid map is defined over a continuous space, and it can be discretized with, e.g., a grid approximation. The size of the map can be reduced to a certain area surrounding the ego vehicle. In order to keep a constant map size while the vehicle is moving, some parts of the map are thrown away and new parts are initiated. Occupancy grid mapping (OGM) is one method for tackling the problem of generating consistent maps from noisy and uncertain data under the assumption that the ego vehicle pose, i.e., position and heading, is known. These maps are very popular in the robotics community, especially for all sorts of autonomous vehicles equipped with laser scanners. Indeed, several of the DARPA urban challenge vehicles used OGMs; see Buehler et al. (2008). This is because they are easy to acquire, and they capture important information for navigation. The OGM was introduced by Elfes (1987), and an early introduction is given by Moravec (1988). To the best of the author's knowledge Borenstein and Koren (1991) were the first to utilize OGM for collision avoidance. Examples of OGM in automotive applications are given in Vu et al. (2007). A solid treatment can be found in the recent textbook by Thrun et al. (2005).

This section begins with a brief introduction to occupancy grid maps, according to Thrun et al. (2005). Using this theory and a sensor with high resolution usually gives a nice looking bird eye's view map. However, since a standard automotive radar is used, producing only a few range and bearing measurements at every time sample, some modifications are introduced as described in the following sections.

5.1 Background

The planar map \mathbf{M} is defined in the world coordinate frame W and is represented by a matrix. An occupancy grid map is partitioned into finitely many grid cells

$$\mathbf{M} = \left\{ \mathbf{m}^{(j)} \right\}_{j=1}^{N_m}. \quad (15.57)$$

The probability of a cell being occupied $p(\mathbf{m}^{(j)})$ is specified by a number ranging from 1 for occupied to 0 for free. The notation $p(\mathbf{m}^{(j)})$ will be used to refer to the probability that a grid cell is occupied. A disadvantage with this design is that it does not allow for dependencies between neighboring cells.

The occupancy grid map was originally developed to primarily be used with measurements from a laser scanner. A laser is often mounted on a rotating shaft, and it generates a range measurement for every angular step of the mechanical shaft, i.e., a bearing angle. This means that the continuously rotating shaft produces many range and bearing measurements during every cycle. The OGM algorithms transform the polar coordinates of the measurements into Cartesian coordinates in a fixed world or map frame. After completing one mechanical measurement cycle, the sensor provides the measurements for use.

The algorithm loops through all cells and increases the occupancy probability $p(\mathbf{m}^{(j)})$ if the cell was occupied according to the measurement $\mathbf{z}_k^{(i)}$. Otherwise, the occupancy value either remains unchanged or is decreased, depending on if the range to the cell is greater or less than the measured range. The latter implies that the laser beam did pass this cell without observing any obstacles. If the measured range is too large or the cell size is too small, it might be necessary to consider the angular spread of the laser beam and increase or decrease the occupancy probability of several cells with respect to the beam width.

The map is assumed to be static, i.e., it does not change during sensing. In this section, the map estimation problem is solved with a binary Bayes filter, of which OGM is one example. In this case, the estimation problem is solved with the binary Bayes filter, of which OGM is one example. The state can either be free $\mathbf{m}^{(j)} = 0$ or occupied $\mathbf{m}^{(j)} = 1$. A standard technique to avoid numerical instabilities for probabilities close to 0 and to avoid truncation problems close to 0 and 1 is to use the log odds representation of occupancy

$$\ell_{j,k} = \log \frac{p(\mathbf{m}^{(j)} | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k})}{1 - p(\mathbf{m}^{(j)} | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k})}, \quad (15.58)$$

or put in words, the odds of a state is defined as the ratio of the probability of this event $p(\mathbf{m}^{(j)} | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k})$ divided by the probability of its complement $1 - p(\mathbf{m}^{(j)} | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k})$. The probabilities are easily recovered using

$$p(\mathbf{m}^{(j)} | \mathbf{Z}_{1:k}, \mathbf{x}_{E,1:k}) = 1 - \frac{1}{1 + \exp \ell_{j,k}}. \quad (15.59)$$

Note that the filter uses the inverse measurement model $p(\mathbf{m} | \mathbf{z}, \mathbf{x})$. Using Bayes' rule it can be shown that the binary Bayes filter in log odds form is

$$\ell_{j,k} = \ell_{j,k-1} + \log \frac{p(\mathbf{m}^{(j)} | \mathbf{Z}_k, \mathbf{x}_{E,k})}{1 - p(\mathbf{m}^{(j)} | \mathbf{Z}_k, \mathbf{x}_{E,k})} - \log \frac{p(\mathbf{m}^{(j)})}{1 - p(\mathbf{m}^{(j)})}, \quad (15.60)$$

where $p(\mathbf{m}^{(j)})$ represents the prior probability. The log odds ratio of the prior before processing any measurements is defined as

$$\ell_{j,0} = \log \frac{p(\mathbf{m}^{(j)})}{1 - p(\mathbf{m}^{(j)})}. \quad (15.61)$$

Typically, $p(\mathbf{m}^{(j)}) = 0.5$ is assumed, since before having measurements, nothing is known about the surrounding environment. This value yields $\ell_0 = 0$.

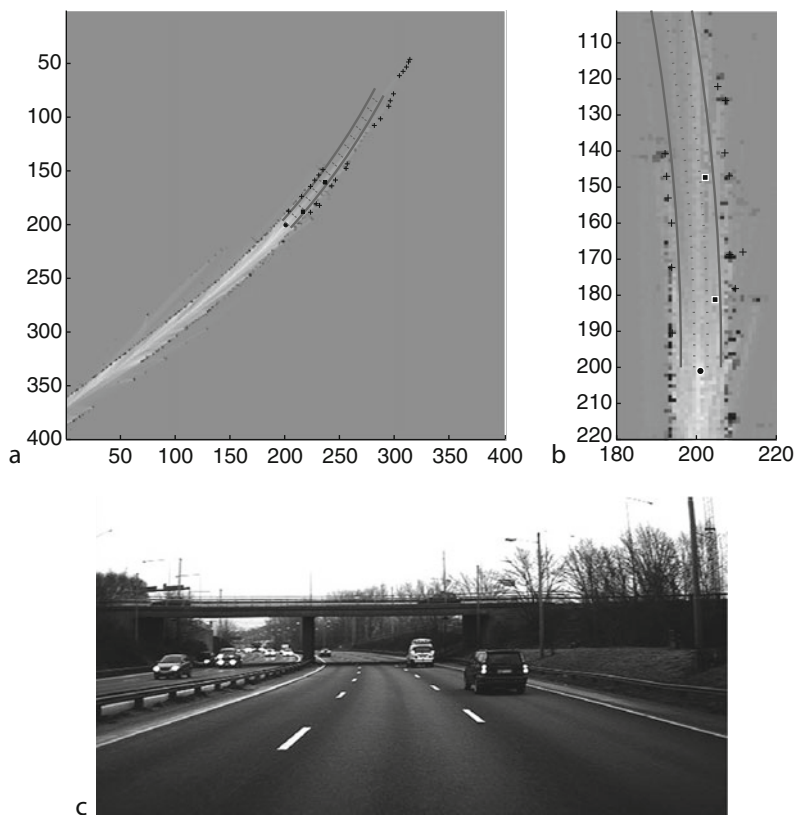
5.2 OGM with Radar Measurements

The radar system provides range and bearing measurements for observed targets at every measurement cycle. The main difference to a laser is that there is not one range measurement for every angular position of the moving sensor. The number of observations depends on the environment. In general, there are much fewer observations compared to a laser sensor. There is also a limit on the number of objects transmitted by the radar equipment on the CAN-bus, and a proprietary selection is performed in the radar. Moving objects, which are distinguished by measurements of the Doppler shift, are prioritized and more likely to be transmitted than stationary objects. Furthermore, it is assumed that the opening angle of the radar beam is small compared to the grid cell size. With these, the OGM algorithm is changed to loop through the measurements instead of the cells in order to decrease the computational load. A radar's angular uncertainty is usually larger than its range uncertainty. When transforming the polar coordinates of the radar measurements into the Cartesian coordinates of the map, the uncertainties can either be transformed in the same manner or it can simply be assumed that the uncertainty increases with the range.

5.3 Experiments and Results

► [Figure 15.6a](#) shows an OGM example of a highway situation. The ego vehicle's camera view is shown in ► [Fig. 15.6c](#). The size of the OGM is 401×401 m, with the ego vehicle in the middle cell. Each cell represents a 1×1 m². The gray-level in the occupancy map indicates the probability of occupancy $p(\mathbf{MIZ}_{1:k} | \mathbf{x}_{E,1:k})$, the darker the grid cell, the more likely it is to be occupied. The map shows all major structural elements as they are visible at the height of the radar. This is a problem if the road is undulated and especially if the radar observes obstacles over and behind the guardrail. In this case, the occupancy probability of a cell might be decreased, even though it was previously believed to be occupied, since the cell is between the ego vehicle and the new observation. The impact of this problem can be reduced by tuning the filter well.

It is clearly visible in ► [Fig. 15.6a](#) that the left border is sharper than the right. The only obstacle on the left side is the guardrail, which gives rise to the sharp edge, whereas on the right side, there are several obstacles behind the guardrail, which also cause reflections, e.g., noise barrier and vegetation. A closer look at ► [Fig. 15.6b](#) reveals that there is no black line of occupied cells representing the guardrail as expected. Instead, there is a region with mixed probability of occupancy and, after about 5 m, the gray region with initial valued cells tells us that nothing is known about these cells.



■ Fig. 15.6

The filled *circle* at position (201, 201) in the occupancy grid map in (a) is the ego vehicle, the *+* are the radar observations obtained at this time sample, the *black squares* are the two leading vehicles that are currently tracked. (b) shows a zoom of the OGM in front of the ego vehicle. The *gray-level* in the figure indicates the probability of occupancy, the darker the grid cell, the more likely it is to be occupied. The shape of the road is given as solid and dashed lines, calculated as described in ▶ Sect. 4. The camera view from the ego vehicle is shown in Fig. (c); the concrete walls, the guardrail, and the pillar of the bridge are interesting landmarks. Furthermore, the two tracked leading vehicles are clearly visible in the *right* lane

6 Intensity-Based Map

The bin-occupancy filter, which is described in Erdinc et al. (2009), aims at estimating the probability of a target being in a given point. The approach is derived via a discretized state-space model of the surveillance region, where each grid cell (denoted bin in this approach) can or may not contain a target. One of the important assumptions is that the bins are sufficiently small so that each bin is occupied by maximum one target. In the

limiting case, when the volume of the bins $|v|$ tends to zero, it is possible to define the bin-occupancy density

$$D_{k|k} \triangleq \lim_{|v| \rightarrow 0} \frac{\Pr(\mathbf{m}_k^{(j)} = 1 | \mathbf{Z}_{1:k})}{|v|}, \quad (15.62)$$

where $\Pr(\mathbf{m}_k^{(j)} = 1 | \mathbf{Z}_{1:k})$ is the probability that bin j is occupied by one target. The continuous form of the bin-occupancy filter prediction and update equations are the same as the probability hypothesis density (PHD) filter equations (Erdinc et al. 2009). Furthermore, the PHD is the first moment density or *intensity density* in point process theory (see e.g., Mahler (2007)), and a physical interpretation is given in Daley and Vere-Jones (2003) as the probability that one target is located in the infinitesimal region $(\mathbf{x}, \mathbf{x} + d\mathbf{x})$ of the state space, divided by $d\mathbf{x}$. The continuous form of the physical bin model leads us to a continuous location-based map, which we denote *intensity-based map*, and intend to estimate with the PHD filter.

The bin-occupancy filter or the PHD filter was developed for target tracking of point sources; however, the aim in this section is to create a probabilistic location-based map of the surroundings of a moving vehicle. One of the main differences between standard target tracking problems and the building of a location-based map is that many objects, such as guardrails or walls, are typically not point targets but extended targets (Mahler 2007, Gilholm and Salmond 2005). Furthermore, there is no interest in estimating the number of objects in the map, and there is also no interest in keeping track of specific objects. Nevertheless, the bin-occupancy filter attempts to answer the important question: “Is there an object (target) at a given point?” Erdinc et al. (2009) pose the following assumptions for the bin-occupancy filter:

1. The bins are sufficiently small so that each bin is occupied by at most one target.
2. One target gives rise to only one measurement.
3. Each target generates measurements independently.
4. False alarms are independent of target originated measurements.
5. False alarms are Poisson distributed.

Here, only point 2 needs some extra treatment if the aim of the algorithm is mapping and not target tracking. It can be argued that the measurements of point sources belong to extended objects and that the aim is to create a map of those point sources. Also for mapping purposes, the assumption that there will not be two measurements from the same point at the same time is justified. The relation described is modeled by a likelihood function $p(\mathbf{Z}_k | \mathbf{M}_{k|k})$, which maps the Cartesian map to polar point measurements.

So far in this section, the discussion has been quite general, and the PHD or the intensity has only been considered as a surface over the surveillance region. The first practical algorithms to realize the PHD filter prediction and measurement update equations were based on the particle filter (see, e.g., Vo et al. (2003), Sidenblad (2003) where the PHD is approximated by a large set of random samples (particles)). A Gaussian mixture approximation of the PHD (GM-PHD) was proposed by Vo and Ma (2006).

The mixture is represented by a sum of weighted Gaussian components and, in particular, the mean and covariance of those components are propagated by the Kalman filter. In this work, we represent the intensity by a Gaussian mixture since the parametrization and derivation is simpler than for a particle filter-based solution. The modeling of the intensity through a number of Gaussian components also makes it simpler to account for structures in the map.

The GM-PHD filter estimates the posterior intensity, denoted $D_{k|k}$, as a mixture of Gaussian densities as,

$$D_{k|k} = \sum_{i=1}^{J_{k|k}} w_{k|k}^{(i)} N\left(m_{k|k}^{(i)}, P_{k|k}^{(i)}\right), \quad (15.63)$$

where $J_{k|k}$ is the number of Gaussian components and $w_{k|k}^{(i)}$ is the expected number of point sources covered by the density $N\left(m_{k|k}^{(i)}, P_{k|k}^{(i)}\right)$. In Lundquist et al. (2011a), it is shown how the intensity is estimated with the GM-PHD filter. The Gaussian components are parametrized by a mean $m_{k|k}^{(i)}$ and a covariance $P_{k|k}^{(i)}$, which are expressed in a planar Cartesian coordinate frame, according to

$$m_k^{(i)} = \begin{bmatrix} x_k^{(i)} & y_k^{(i)} \end{bmatrix}^T. \quad (15.64)$$

The aim of the mapping algorithm is to estimate the posterior density (► 15.3). The considered intensity-based map is continuous over the surveillance region; thus, for the number of elements in (► 15.1), it holds that $N_m \rightarrow \infty$. Furthermore, the intensity is a summary statistic of the map according to (see e.g., Mahler [2003])

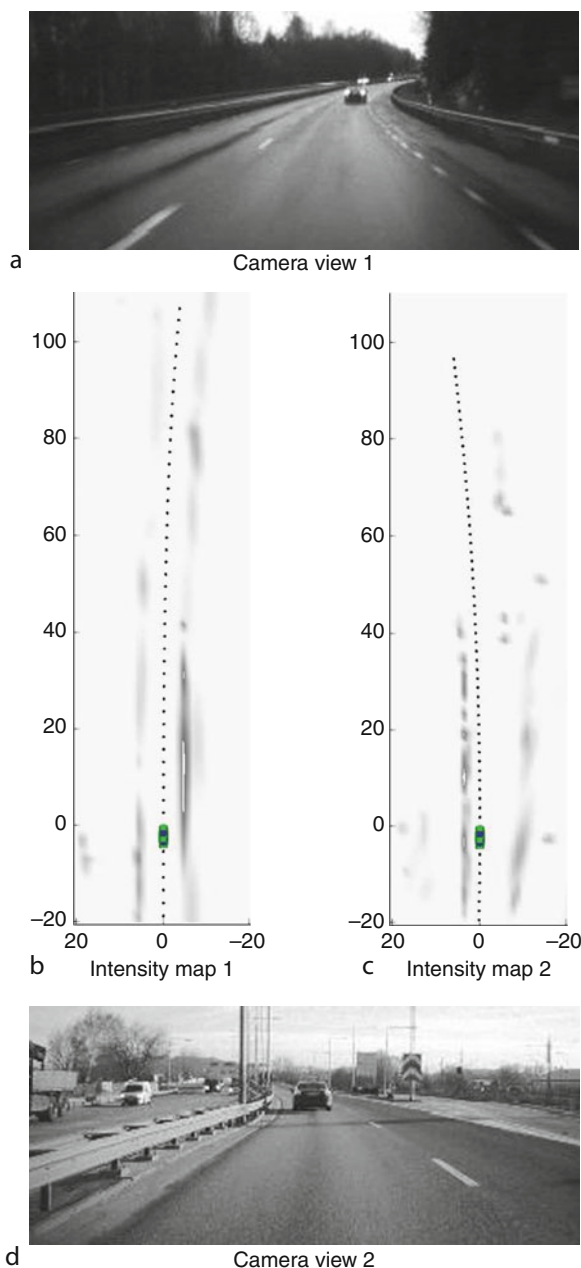
$$p(\mathbf{M}_k | \mathbf{Z}_{1:k}) \sim p(\mathbf{M}_k; D_{k|k}), \quad (15.65)$$

and the estimated intensity $D_{k|k}$ is parametrized by

$$\mu_k^{(i)} \triangleq \left\{ w_k^{(i)}, m_k^{(i)}, P_k^{(i)} \right\} \quad (15.66)$$

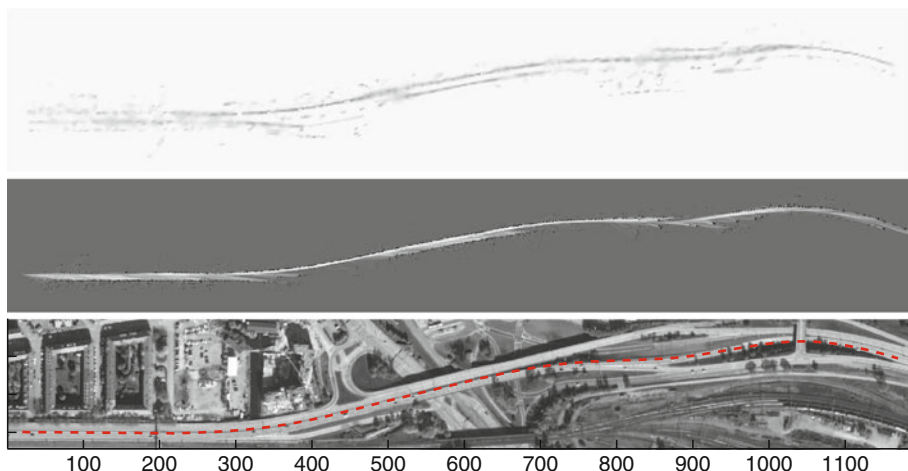
of the Gaussian sum (► 15.63). The intensity-based map is a multimodal surface with peaks around areas with many sensor reflections or point sources. It is worth observing that the map \mathbf{M} is described by a location-based function (► 15.63), with feature-based parametrization (► 15.66).

Experiments were conducted with a prototype passenger car. One example of the estimated intensity at a freeway traffic scenario is shown as a bird's eye view in ► Fig. 15.7b. Darker regions illustrate higher concentrations of point sources, which in this figure stem from the guardrails to the left and the right of the road. As expected, the path of the ego vehicle, indicated by the black dots, is in between the two regions of higher object concentration. The driver's view is shown in ► Fig. 15.7a.



■ Fig. 15.7

The image in (a) shows the driver's view of the intensity map in (b), and the image in (d) is the driver's view of the intensity map in (c). The darker the areas in the intensity map, the higher the concentration of objects. The driver's path is illustrated with *black* dots and may be used as a reference. Note that snap shot in (c) and (d) is obtained only some meters after the situation shown in ► Fig. 15.1



■ Fig. 15.8

The *top* figure shows the intensity-based map obtained from radar measurements collected on a freeway. The OGM in the middle figure serves as a comparison of an existing algorithm. The *bottom* figure is a flight photo used as ground truth, where the driven trajectory is illustrated with a dashed line (© Lantmäteriet Gävle 2010. Medgivande I 2011/0100.

Reprinted with permission). Note that the drivers view at 295 m is shown in ◀ Fig. 15.7d, and about the same position is also shown in ▶ Figs. 15.1 and ▶ 15.5

A second example is shown in ▶ Fig. 15.7c and ▶ d. Here, the freeway exit is clearly visible in the intensity map, which shows that the proposed method to create maps is very conformable.

The Gaussian components are generally removed from the filter when the vehicle passed those parts of the map. However, to give a more comprehensive overview, these components are stored, and the resulting intensity-based map is shown together with an occupancy grid map (OGM) and a flight photo in ▶ Fig. 15.8. The top figure is the map produced as described in this section. The OGM, described in ▶ Sect. 5, is based on the same data set and used as a comparison. The gray-level of the OGM indicates the probability of occupancy: the darker the grid cell, the more likely it is to be occupied. As seen in the figure, the road edges are not modeled as distinct with the OGM. The OGM representation of the map is not very efficient since huge parts of the map are gray indicating that nothing is known about these areas. An OGM matrix with often more than 10,000 elements must be updated and communicated to other safety functions of a car at each time step. The compact representation is an advantage of the intensity-based map. Each Gaussian component is parametrized with 7 scalar values according to (▶ 15.66). Since most maps are modeled with 10–30 components, it summarizes to around 70–210 scalar values, which easily can be sent on the vehicles CAN-bus to other safety functions. Finally, the bottom photo is a very accurate flight photo (obtained from the Swedish mapping, cadastral, and land registration authority), which can be used as ground truth to visualize the quality of the intensity-based map.

7 Conclusion

The use of radar, laser, and camera for situation awareness is gaining popularity in automotive safety applications. In this chapter, it has been shown how sensor data perceived from the ego vehicle is used to estimate a map describing the local surroundings of a vehicle. The map may be modeled in various different ways, of which four major approaches have been described. In a feature-based map, each element of the map specifies the properties and location of one object. This can either be a point source in the space; or it can be an extended object such as the position and shape of the lane or the road edges. Furthermore, in a location-based map, the index of each element corresponds to a location and the value of the map element describes the property in that position. One example is the occupancy grid map, which is defined over a continuous space but discretized with a grid approximation. Another example is the intensity-based map, which is a continuous approximation, describing the density of objects in the map. The four approaches presented in this chapter have all been evaluated on real data from both freeways and rural roads in Sweden.

The current accuracy of GPS receivers is acceptable only for route guidance, where the provided global position is sufficient. For automotive active safety systems, the local position of the ego vehicle with respect to its surroundings is more important. The estimated maps, described in this chapter, can be used to increase the localization accuracy of the ego vehicle. Furthermore, the maps may be used to derive a collision free trajectory for the vehicle.

Acknowledgments

The authors would like to thank the SEnsor Fusion for Safety (SEFS) project within the Intelligent Vehicle Safety Systems (IVSS) program, the strategic research center MOVIII, funded by the Swedish Foundation for Strategic Research (SSF) and CADICS, a Linneaus Center funded by the Swedish Research Council for financial support.

References

- Björck Å (1996) Numerical methods for least squares problems. SIAM, Philadelphia
- Blackman SS, Popoli R (1999) Design and Analysis of Modern Tracking Systems. Artech House, Norwood
- Borenstein J, Koren Y (1991) The vector field histogram-fast obstacle avoidance for mobile robots. IEEE Trans Robot Automation 7(3):278–288
- Buehler M, Iagnemma K, Singh S (eds) (2008) Special issue on the 2007 DARPA urban challenge, Part I-III. J Field Rob 25:8–10
- Caveney D (2010) Cooperative vehicular safety applications. IEEE Control Syst Mag 30(4):38–53
- Civera J, Davison A, Montiel J (2008) Inverse depth parametrization for monocular SLAM. IEEE Trans Rob 24(5):932–945
- Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes, vol 1, 2nd edn, Elementary theory and method. Springer, New York
- Davison AJ, Reid I, Molton N, Strasse O (2007) MonoSLAM: Real-time single camera SLAM. IEEE Trans Pattern Anal Mach Intell 29(6):1052–1067
- Dickmanns E (1988) Dynamic computer vision for mobile robot control. In: Proceedings of the

- international symposium on industrial robots, Sydney
- Dickmanns ED (2007) Dynamic vision for perception and control of motion. Springer, London
- Dickmanns ED, Mysliwetz BD (1992) Recursive 3-D road and relative ego-state recognition. *IEEE Trans Pattern Anal Mach Intell* 14(2):199–213
- Diversi R, Guidorzi R, Soverini U (2005) Kalman filtering in extended noise environments. *IEEE Trans Autom Control* 50(9):1396–1402
- Duda RO, Hart PE (1972) Use of the Hough transformation to detect lines and curves in pictures. *Commun ACM* 15(1):11–15
- Eidehall A, Pohl J, Gustafsson F, Ekmark J (2007) Toward autonomous collision avoidance by steering. *IEEE Trans Intell Transp Syst* 8(1):84–94
- Elfes A (1987) Sonar-based real-world mapping and navigation. *IEEE J Robot Automation* 3(3):249–265
- Erdinc O, Willett P, Bar-Shalom Y (2009) The bin-occupancy filter and its connection to the PHD filters. *IEEE Trans Signal Process* 57(11):4232–4246
- Gilholm K, Salmond D (2005) Spatial distribution model for tracking extended objects. *IEE Proc Radar Sonar Navigation* 152(5):364–371
- Gill PE, Murray W, Saunders MA, Wright MH (1991) Inertia-controlling methods for general quadratic programming. *SIAM Rev* 33(1):1–36
- Goldfarb D, Idnani A (1983) A numerically stable dual method for solving strictly convex quadratic programs. *Math Program* 27(1):1–33
- Gordon NJ, Salmond DJ, Smith AFM (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc Radar Signal Process* 140(5):107–113
- Guiducci A (1999) Parametric model of the perspective projection of a road with applications to lane keeping and 3D road reconstruction. *Comput Vis Image Underst* 73(3):414–427
- Guiducci A (2000) Camera calibration for road applications. *Comput Vis Image Underst* 79(2):250–266
- Harris C, Stephens M (1988) A combined corner and edge detector. In: *Proceedings of the Alvey vision conference*, Manchester, pp 147–151
- Hough PVC (1962) A method and means for recognizing complex patterns. US Patent 3,069,654
- Jung CR, Kelber CR (2005) Lane following and lane departure using a linear-parabolic model. *Image Vis Comput* 23(13):1192–1202
- Kim ZW (2008) Robust lane detection and tracking in challenging scenarios. *IEEE Trans Intell Transp Syst* 9(1):16–26
- Lee JW (2002) A machine vision system for lane-departure detection. *Comput Vis Image Underst* 86(1):52–78
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
- Lundquist C, Schön TB (2011) Joint ego-motion and road geometry estimation. *Inf Fusion* 12(4):253–263
- Lundquist C, Hammarstrand L, Gustafsson F (2011a) Road intensity based mapping using radar measurements with a probability hypothesis density filter. *IEEE Trans Signal Process* 59(4):1397–1408
- Lundquist C, Orguner U, Gustafsson F (2011b) Extended target tracking using polynomials with applications to road-map estimation. *IEEE Trans Signal Process* 59(1):15–26
- Mahler RPS (2003) Multitarget Bayes filtering via first-order multitarget moments. *IEEE Trans Aerosp Electron Syst* 39(4):1152–1178
- Mahler RPS (2007) Statistical multisource-multitarget information fusion. Artech House, Boston
- Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide baseline stereo from maximally stable extremal regions. *Image Vis Comput* 22(10):731–767
- McCall JC, Trivedi MM (2006) Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation. *IEEE Trans Intell Transp Syst* 7(1):20–37
- Moravec H (1988) Sensor fusion in certainty grids for mobile robots. *AI Mag* 9(2):61–74
- Nistér D, Naroditsky O, Bergen J (2006) Visual odometry for ground vehicle applications. *J Field Rob* 23(1):3–20
- Powell M (1985) On the quadratic programming algorithm of Goldfarb and idnani. *Math Program Study* 25(1):46–61
- Rohling H, Meinecke MM (2001) Waveform design principles for automotive radar systems. In: *Proceedings on CIE international conference on radar*, Beijing, pp 1–4
- Rohling H, Möller C (2008) Radar waveform for automotive radar systems and applications. In: *IEEE radar conference*, Rome, pp 1–4
- Sidenbladh H (2003) Multi-target particle filtering for the probability hypothesis density.

- In: Proceedings of the international conference on information fusion, Cairns, vol 2, pp 800–806
- Söderström T (2007) Survey paper: errors-in-variables methods in system identification. *Automatica* 43(6):939–958
- Szeliski R (2010) *Computer vision: algorithms and applications*. Springer, New York
- Thrun S, Burgard W, Fox D (2005) *Probabilistic robotics*. The MIT Press, Cambridge
- Vo B-N, Ma W-K (2006) The Gaussian mixture probability hypothesis density filter. *IEEE Trans Signal Process* 54(11):4091–4104
- Vo B-N, Singh S, Doucet A (2003) Random finite sets and sequential monte carlo methods in multi-target tracking. In: Proceedings of the international radar conference, Adelaide, pp 486–491
- Vu TD, Aycard O, Appenrodt N (2007) Online localization and mapping with moving object tracking in dynamic outdoor environments. In: Proceedings of the IEEE intelligent vehicles symposium. Istanbul, pp 190–195
- Wang Y, Bai L, Fairhurst M (2008) Robust road modeling and tracking using condensation. *IEEE Trans Intell Transp Syst* 9(4):570–579
- Waxman A, LeMoigne J, Davis L, Srinivasan B, Kushner T, Liang E, Siddalingaiah T (1987) A visual navigation system for autonomous land vehicles. *IEEE J Robot Automation* 3(2):124–141
- Zhou Y, Xu R, Hu X, Ye Q (2006) A robust lane detection and tracking method based on computer vision. *Meas Sci Technol* 17(4):736–745

16 Navigation and Tracking of Road-Bound Vehicles Using Map Support

Fredrik Gustafsson · Umut Orguner · Thomas B. Schön · Per Skoglar · Rickard Karlsson

Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, SE, Sweden

1	<i>Introduction</i>	399
2	<i>State Estimation and Representation</i>	402
2.1	Nonlinear Filtering	402
2.1.1	Kalman Filter Variants	403
2.1.2	Point Mass and Particle Filter Variants	404
2.1.3	Finite State Space Models	405
2.2	Introductory Illustrations	405
3	<i>Basic Motion Models</i>	407
3.1	Dead-Reckoning Model	408
3.2	A Complete Matlab Algorithm	409
3.3	Tracking Model	411
3.4	Manifold Model	411
4	<i>Map Handling</i>	412
4.1	The Shape Format	412
4.2	Computational Issues Related to the Map	413
5	<i>Navigation Applications</i>	415
5.1	Odometric Approach	415
5.2	Odometric Approach with Parameter Adaptation	417
5.3	Inertial Measurement Support	417

6 *Tracking Approaches* 421

6.1 Radar Support 421

6.2 Wireless Radio Network Support 422

6.3 Sensor Network Support 425

6.4 Vision Support 428

7 *Conclusions* 431





Abstract: The performance of all navigation and tracking algorithms for road-bound vehicles can be improved by utilizing the trajectory constraint imposed from the road network. We refer to this approach as road-assisted navigation and tracking. Further, we refer to the process of incorporating the road constraint into the standard filter algorithms by dynamic map matching. Basically, dynamic map matching can be done in three different ways: (1) as a virtual measurement, (2) as a state noise constraint, or (3) as a manifold estimation problem where the state space is reduced. Besides this basic choice of approach, we survey the field from various perspectives: which filter that is applied, which dynamic model that is used to describe the motion of the vehicle, and which sensors that are used and their corresponding sensor models. Various applications using real data are presented as illustrations.

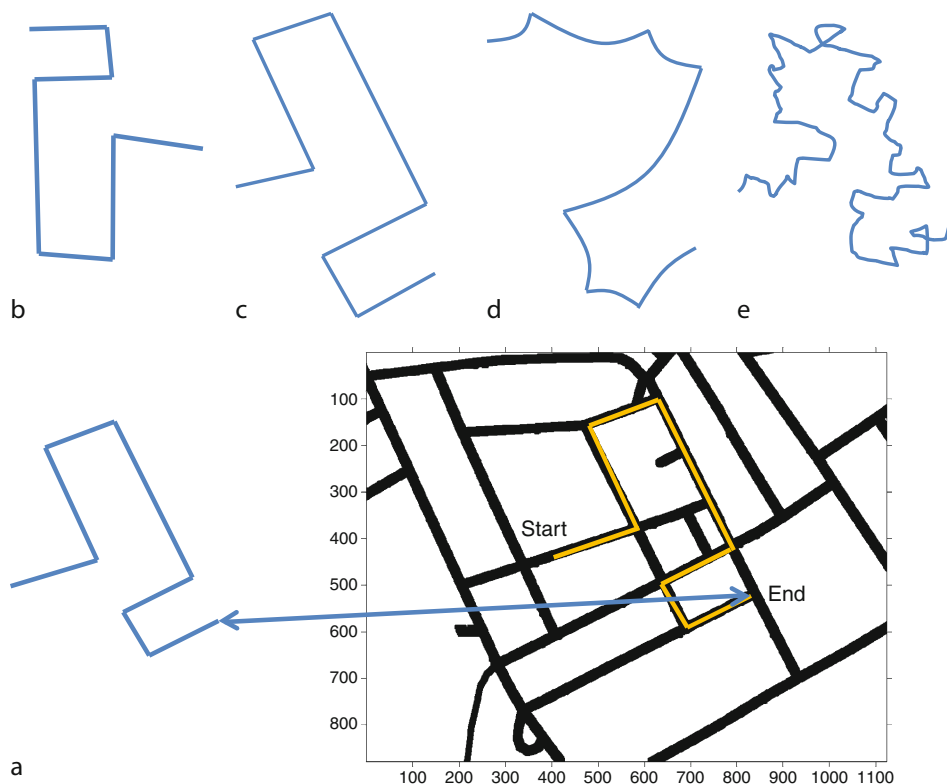
1 Introduction

There is a variety of localization, navigation, and tracking applications that can be improved by restricting the location to be on roads marked on an available map. Essentially, the different applications are distinguished by what sensor combination that is used, and if the computations are done in the vehicle (navigation) or in the infrastructure (tracking). They all have in common that the on-road assumption greatly improves the accuracy, and that even quite poor signal to noise ratio of the sensors is sufficient to get a fairly good navigation performance.

The classical method to improve localization performance is map matching. Here, the position estimate computed from the sensors is mapped to the closest point in the road network. This is an appropriate method for presentation purposes, but it suffers from two problems. The first one is that it does not take the topography of the map into account, which implies that the localization can jump from one road to another due to quantization effects. The second one is that the motion dynamics of the vehicle is not combined with the map information in an optimal way. The purpose of this chapter is to survey different methods to what we will refer to as dynamic map matching.

Dynamic map matching combines a motion model, sensor models, and the road network model in a nonlinear filter, taking uncertainties in all these three kind of models into account.

The problem boils down to fitting a distorted and noisy trajectory to the road network.  *Figure 16.1* illustrates the principle and the basic problems considered one by one. In reality, several of these effects are combined. A typical example is navigation based on odometry, where the wheel speeds are integrated into a trajectory. The unknown absolute radius of the wheels implies a scaling error as in  *Fig. 16.1c*, the relative radii difference gives a bias in the yaw rate corresponding to  *Fig. 16.1d*, and the absolute course is not observed as illustrated in  *Fig. 16.1b*. Furthermore, the computed trajectory is uncertain due to noisy wheel speed measurements.



■ Fig. 16.1

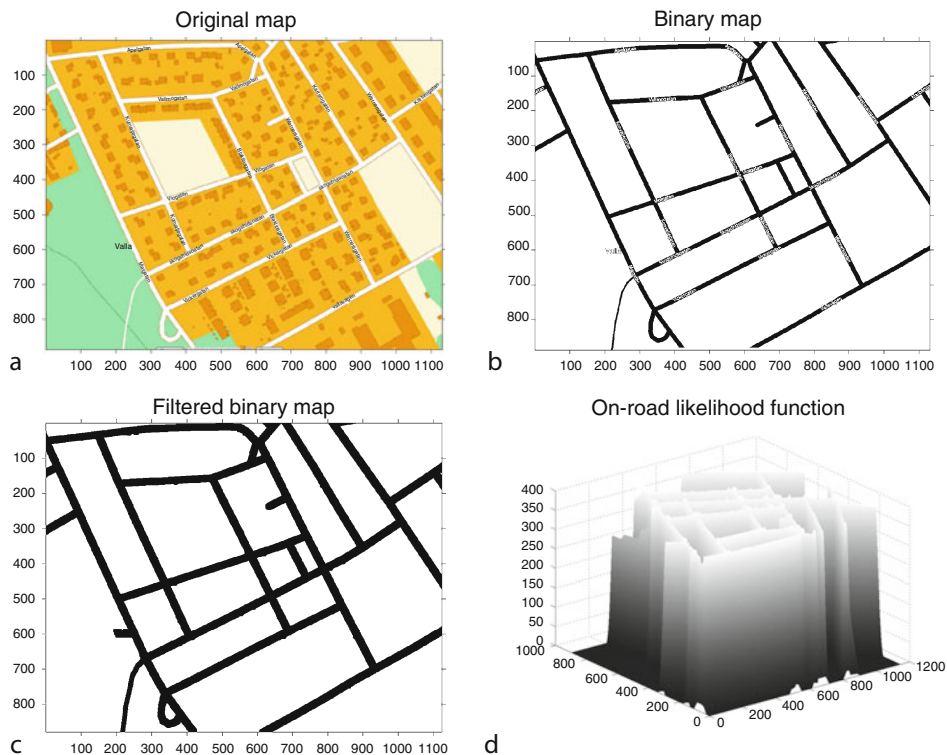
The key idea in dynamic map matching is to fit an observed trajectory to the road network. (a) Undistorted trajectory. (b) Undistorted trajectory with random rotation. (c) Trajectory based on biased speed. (d) Trajectory based on biased yaw rate. (e) Trajectory with random noise

A generic nonlinear filter for navigation consists of the following main steps:

- *Time update or prediction:* Use a motion model to predict where the vehicle will be when the next measurement arrives.
- *Measurement update or correction:* Use the current measurement and a sensor model to update the information about the current location.

In a Bayesian framework, the information is represented by the posterior distribution given all available measurements. The process of computing the Bayesian posterior distribution is called filtering.

➤ *Figure 16.2* illustrates how a standard map can be converted to a likelihood function for the position. Positions on roads get the highest likelihood, and the likelihood quickly decays as the distance to the closest road increases. A small offset can be added to the likelihood function to allow for off-road driving, for instance on unmapped roads and parking areas.



■ Fig. 16.2

(a) Original map. (b) Binary map, where the black areas corresponding to streets are mapped to one, and all other pixels are set to zero (white color). (c) The local maxima over a 4×4 region is computed to remove text information. (d) The resulting map is low-pass filtered to allow for small deviations from the road borders, which yields a smooth likelihood function for on-road vehicles. See Listing 1 for complete Matlab code

There are three main principles for incorporating the on-road constraint:

1. As a *virtual measurement* using a road-tailored likelihood function, see ► Fig. 16.2d, where predicted positions outside the road network are considered unlikely.
2. As a *state noise constraint* in the prediction step, so the predicted position is mostly on a road. Here, the likelihood in ► Fig. 16.2d is instrumental.
3. As *manifold filtering*, where the location is represented as the position along a road segment. This uses the topography of the map in a natural way.

As indicated explicitly in the first two cases, it is in practice necessary to allow the vehicle to temporary leave the road network to allow for off-road driving and unmapped roads such as in parking areas and houses. This can be solved by having two modes in the filter, one on-road mode and one off-road mode, respectively.

There are two classes of problems with different support sensor options:

Navigation as driver information or for driver assistance systems as a GPS backup/support. Support sensors include wheel speed, inertial sensors, and visual odometry using visual landmarks.

Tracking for surveillance or traffic monitoring. Support sensors include imagery sensors, radar, and sensor networks with microphones, geophones, or magnetometers.

Tracking and navigation are in some sense dual problems, and the difference disappears in a cooperative setting where all equipment exchange information. Localization in cellular systems is one example, where network-centric (tracking) and user-centric (navigation) solutions exist, which are more or less similar. The main difference lies in where the algorithm is implemented, the algorithm itself can be the same. Here, we define tracking as all approaches that require infrastructure with communication ability (to exclude visual markers and passive radio beacons). In the sequel, we often use the term navigation for both classes of problems.

The outline is as follows. 🔗 [Section 2](#) surveys the different nonlinear filters that have been used for road-assisted navigation and tracking, with some illustrations from applications to illustrate the different concepts. 🔗 [Section 3](#) presents three fundamental and basic motion models and provides concrete code. 🔗 [Section 4](#) discusses the data format used in the map and explains the mathematical map matching operation. 🔗 [Section 5](#) summarizes some navigation applications, while 🔗 [Sect. 6](#) overviews some tracking applications from our earlier research publications.

2 State Estimation and Representation

2.1 Nonlinear Filtering

We consider a general nonlinear motion model for the road-bound target with state x_k , position dependent measurement y_k , input signal u_k , process noise w_k , and measurement noise e_k :

$$x_{k+1} = f(x_k, u_k, v_k), \quad (16.1a)$$

$$y_k = h(x_k, u_k, e_k). \quad (16.1b)$$

The state includes at least position (X_k, Y_k) and heading (or course) ψ_k , and possibly derivatives of these and further parameters and states relevant in describing the motion.

Nonlinear filtering is the branch of statistical signal processing concerned with recursively estimating the state x_k in (🔗 [16.1a](#)) based on the measurements up to time k , $y_{1:k} \triangleq \{y_1, \dots, y_k\}$ from sample 1 to k . The most general problem it solves is to compute the Bayesian conditional posterior density $p(x_k | y_{1:k})$.

There are several algorithms and representations for computing the posterior density:

- Kalman filter variants: The state is represented with a Gaussian distribution.
- Kalman filter banks based on multiple model approaches: The state is represented with a mixture of Gaussian distributions, each Gaussian mode having associated weights.
- Point mass and particle filters: The state is represented with a set of grid points or samples with an associated weight.
- Marginalized, or Rao-Blackwellized, particle filters: The state is represented with a number of trajectories over the road network, each one having an associated weight and Gaussian distribution for the other state variables than position.
- Finite state space models: The trajectory is represented by discrete probabilities for each combination of possible turns in the road network junctions.

These different filters are briefly introduced below.

2.1.1 Kalman Filter Variants

The Kalman filter (KF) (Kalman 1960) solves the filtering problem in case the model (16.1a, b) is linear and Gaussian. The solution involves propagating the mean $\hat{x}_{k|k}$ and the covariance $P_{k|k}$ for the posterior Gaussian distribution

$$p(x_k | y_{1:k}) = \mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k}). \quad (16.2)$$

The extended Kalman filter (EKF) (Schmidt 1966) and the unscented Kalman filter (UKF) (Julier et al. 1995) approximate the posterior at each step with a Gaussian density according to (16.2).

The road constraints imply a kind of information that normally leads to a multimodal posterior density (the target can be on either this road or that road, etc.). The approximation in using (16.2) inevitably destroys this information. A completely different approach to nonlinear filtering is based on approximating the posterior $p(x_k | y_{1:k})$ numerically. One straightforward extension is to assign one KF to each hypothesis in a multiple model (MM) framework, which leads to a Kalman filter bank (KFB) technique with a Gaussian mixture posterior approximation (Sorenson and Alspach 1972),

$$p(x_k | y_{1:k}) \approx \sum_{i=1}^N w_k^i \mathcal{N}(x_k; \hat{x}_{k|k}^i, P_{k|k}^i). \quad (16.3)$$

The mixture probabilities w_k^i are all positive and sum to one. Each Gaussian distribution can be interpreted as a conditional distribution given one specific hypothesis about how the driven path matches the road-map topography, where the observed sequence of turns fits the map in different ways. The Manhattan problem is used to illustrate the combinatoric explosion of hypotheses that are possible in a regular road pattern. There are two conceptually different ways to limit the number of hypotheses: pruning where unlikely hypotheses are thrown away, and merging where similar hypotheses which

end up at the same position are merged into one. The interacting multiple model (IMM) (Blom and Bar-Shalom 1988) algorithm is a popular choice for the latter approach. Since the number of modes varies depending on excitation, compare with ④ Fig. 16.9, the concept of variable structure (VS) has been adopted in the literature, see for instance Li and Bar-Shalom (1996).

2.1.2 Point Mass and Particle Filter Variants

The class of point mass filters (PMF) (Kramer and Sorenson 1988) represents the state space using a regular grid of size N , where the grid points and the related weights (x_i, w_k^i) are used as a representation of the posterior. Different basis functions have been suggested, the simplest one being an impulse at each grid, when the posterior approximation can be written

$$p(x_k|y_{1:k}) \approx \sum_{i=1}^N w_k^i \delta(x_k - x_k^i), \quad (16.4)$$

where $\delta(x)$ denotes the Dirac delta function. The particle filter (PF) (Gordon et al. 1993) is the state of the art numerical solution today. It uses a stochastic grid $\{w_k^i, x_k^i\}_{i=1}^N$ that automatically changes at each iteration. Another difference is that it, in its standard form, approximates the trajectory $x_{1:k}$. Otherwise, the representation of the posterior approximation is very similar to (④ 16.4),

$$p(x_{1:k}|y_{1:k}) \approx \sum_{i=1}^N w_k^i \delta(x_{1:k} - x_{1:k}^i). \quad (16.5)$$

One should here note that in many navigation applications the sensor model is only a function of position. With an assumption of additive noise, the sensor model in (④ 16.1b) can in such cases be written

$$y_k = h(X_k, Y_k) + e_k. \quad (16.6)$$

If the state x_k only includes position and velocity $x_k = (X_k, Y_k, \dot{X}_k, \dot{Y}_k)^T$, which is the simplest possible standard model in target tracking, then the marginalized particle filter (MPF, also known as RBPF, the Rao-Blackwellized PF) applies. The basic idea in the MPF is to utilize the structure in the model, so the Kalman filter can be applied to a part of the state vector (the velocity $V_k = (\dot{X}_k, \dot{Y}_k)^T$ in this case) in an optimal way. The resulting posterior approximation is then

$$p(x_{1:k}|y_{1:k}) \approx \sum_{i=1}^N w_k^i \delta(X_{1:k} - X_{1:k}^i) \delta(Y_{1:k} - Y_{1:k}^i) \mathcal{N}(V_k; \hat{V}_{k|k}^i, P_{k|k}^i). \quad (16.7)$$

This means that each particle represents a trajectory in the map, which has an associated Gaussian distributed velocity vector attached to it. The MPF can be applied to many other cases where the motion model contains more states than just position and heading. One of our key messages is that the MPF is well suited for road-assisted navigation and tracking.

2.1.3 Finite State Space Models

All approaches so far have assumed a continuous state space based on the 2D position. A completely different approach is based on a discrete state representing road segments defined by its junctions with other road segments. Let m denote a certain road segment (possibly one-way to indicate the direction of travel). Its end is connected to a number of other road segments n_1, n_2, \dots, n_m . Let the transition probabilities from road m to another road n be defined as

$$\text{Prob}(n|m) = \begin{cases} \pi_{nm} & n = n_1, n_2, \dots, n_m, \\ 0 & \text{otherwise.} \end{cases} \quad (16.8)$$

Then the complete road topology and prior knowledge of driving behavior is summarized in the matrix Π with elements π_{nm} . The discrete hidden Markov model (HMM) theory provides the optimal filter for estimating the road segment sequence. The basic sampling rate depends on an event process, triggered by an external detection mechanism for indicating when a junction is reached. The estimated sequence in the original sampling rate indexed by k can be obtained by repeating the road segment between the junctions,

$$p(m_{1:k}|y_{1:k}) = \sum_{i=1}^N w_k^i \delta(m_{1:k} - m_{1:k}^i). \quad (16.9)$$

In this case, δ is the discrete pulse function.

Assume that the length of road segment i is L^i , and let $l_k^i \in [0, L^i]$ be the driven distance at this road segment. Then, we can form a joint state vector x_k with the continuous state l_k^i , and possibly one or more derivatives of position, for each discrete mode. The two problems can be combined, and the resulting mixture of discrete and continuous states can be expressed as

$$p(m_{1:k}, x_{1:k}|y_{1:k}) = \sum_{i=1}^N w_k^i \delta(m_{1:k} - m_{1:k}^i) \mathcal{N}(x_k; \hat{x}_{k|k}^i, P_{k|k}^i). \quad (16.10)$$

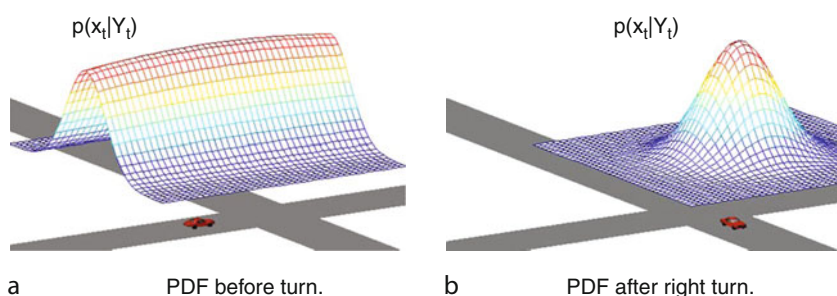
This can be seen as a version of the MPF, where the continuous state can be filtered analytically conditioned on a given sequence of discrete states. The resulting algorithm has a low-dimensional conditional state vector (motion on a manifold) and utilizes the map information in the most accurate way.

2.2 Introductory Illustrations

The Gaussian distribution is in many ways the most convenient representation in a range of applications, but, as already mentioned, a single Gaussian distribution has certain shortcomings for road-assisted navigation. We will here provide a couple of illustrations of posterior distributions, also showing the main principle in road-assisted navigation.

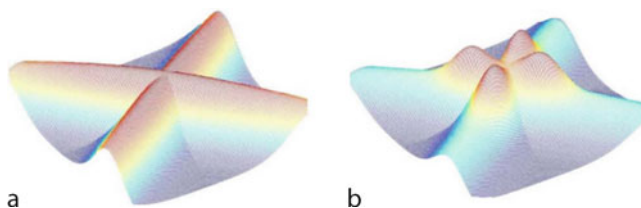
Consider the situation in [Fig. 16.3](#), where a four-way intersection is approached. Suppose that the navigation algorithm has found the correct road segment and direction of travel, but that the information on the driven distance on this segment is uncertain. Then, we get the situation depicted in [Fig. 16.3a](#). Suppose now that the sensors detect a right turn. The posterior distribution in [Fig. 16.3b](#) is then quite informative about the position. This illustrates the information richness in the road map. Here, the Gaussian distribution is a feasible description of both the prior distribution after the prediction step and the posterior after a measurement update from an informative measurement.

The case in [Fig. 16.3](#) assumed prior knowledge about the starting position or the direction of the vehicle. Suppose the prior distribution after the prediction step is instead rather uninformative. This is in [Fig. 16.4a](#) represented with two Gaussian distributions.



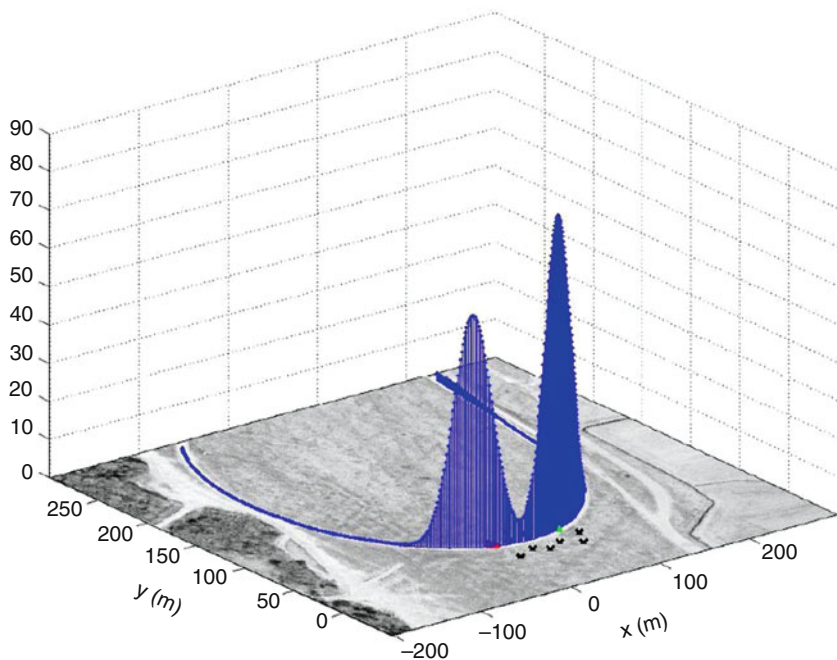
■ Fig. 16.3

(a) The prior of the position close to a four-way intersection when the road segment of the vehicle is known, but its position along the segment is uncertain, can be modeled with a Gaussian distribution. (b) A few meters after a sensed right-hand turn, the posterior distribution becomes very informative (Picture from [Svenzén 2003](#))



■ Fig. 16.4

(a) A non-informative prior of the position close to a four-way intersection can be modeled using a Gaussian mixture with two modes, both centered at the intersection, and each one with a large eigenvalue spread in its covariance matrix along and transversal the road direction. (b) A few meters after a sensed right-hand turn, there are four different possibilities, leading to a Gaussian mixture with four modes (Picture from [Svenzén 2003](#))



■ Fig. 16.5

A Gaussian mixture distribution for modeling the posterior along a road segment, which is marked with a solid line. A microphone network provides the measurements, and each sensor is marked with a cross

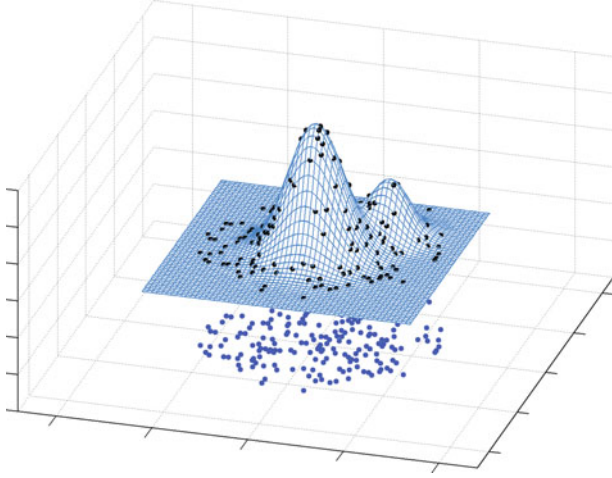
Suppose now again that the sensors detect a right turn. Then, the posterior distribution will have four peaks, each one can be represented with a Gaussian distribution as shown in [Fig. 16.4b](#). If the intersections are regularly spaced (the “Manhattan problem”), it can be hard to resolve such ambiguities.

In manifold filtering, the posterior distribution is constrained to the road network while the filter is operating in the on-road mode. A snapshot illustration is given in [Fig. 16.5](#), where a Gaussian mixture summarizes the information from a sensor network.

Even though (mixtures of) Gaussian distributions are feasible representations of the posterior distribution, a sample-based approximative representation is in many cases even more useful. [Figure 16.6](#) illustrates the key idea: to replace a parametric distribution with samples, or particles.

3 Basic Motion Models

We here describe three specific and simple two-dimensional motion models that are typical for the road-assisted applications.



■ Fig. 16.6

A bimodal Gaussian distribution can as any other distribution be approximated by a set of random samples. This is the idea of the particle filter, and the particle representation of the posterior distribution

3.1 Dead-Reckoning Model

A very instructive and also quite useful motion model is based on a state vector consisting of position (X, Y) and course (yaw angle) ψ . This assumes that there are measurements of yaw rate (derivative of course) $\dot{\psi}$ and speed ϑ on board the platform, in which case the principle of dead-reckoning can be applied.

The dead-reckoning model can be formulated in continuous time using the following equations:

$$x(t) = \begin{pmatrix} X(t) \\ Y(t) \\ \psi(t) \end{pmatrix}, \quad \dot{x}(t) = \begin{pmatrix} \vartheta(t) \cos(\psi(t)) \\ \vartheta(t) \sin(\psi(t)) \\ \dot{\psi}(t) \end{pmatrix} \quad (16.11)$$

A discrete time model for the nonlinear dynamics is given by

$$X(t+T) = X(t) + \frac{2\vartheta(t)}{\dot{\psi}(t)} \sin\left(\frac{\dot{\psi}(t)T}{2}\right) \cos\left(\psi(t) + \frac{\dot{\psi}(t)T}{2}\right) \quad (16.12a)$$

$$\approx X(t) + \vartheta(t)T \cos(\psi(t)),$$

$$Y(t+T) = Y(t) + \frac{2\vartheta(t)}{\dot{\psi}(t)} \sin\left(\frac{\dot{\psi}(t)T}{2}\right) \sin\left(\psi(t) + \frac{\dot{\psi}(t)T}{2}\right) \quad (16.12b)$$

$$\approx Y(t) + \vartheta(t)T \sin(\psi(t)),$$

$$\psi(t+T) = \psi(t) + T\dot{\psi}(t). \quad (16.12c)$$

Finally, plugging in the observed speed $\vartheta^m(t)$ and angular velocity $\dot{\psi}^m(t)$ gives the following dynamic model with process noise $w(t)$

$$X(t+T) = X(t) + \vartheta^m(t)T \cos(\psi(t)) + T \cos(\psi(t))w_{\vartheta}(t), \quad (16.13a)$$

$$Y(t+T) = Y(t) + \dot{\psi}^m(t)T \sin(\psi(t)) + T \sin(\psi(t))w_{\dot{\psi}}(t), \quad (16.13b)$$

$$\psi(t+T) = \psi(t) + T\dot{\psi}^m(t) + T \sin(\psi(t))w_{\dot{\psi}}(t). \quad (16.13c)$$

This model has the following structure

$$x_{k+1} = f(x_k, u_k) + g(x_k, u_k)v_k, \quad u_k = \begin{pmatrix} \vartheta_k^m \\ \dot{\psi}_k^m \end{pmatrix} \quad (16.13d)$$

that fits the particle filter perfectly. Normally, additional support sensors are needed to get observability of the absolute position. This is for instance the case in many robotics applications. However, the road map contains sufficiently rich information in itself.

Note that the speed and the angular velocity measurements are modeled as inputs rather than measurements. This is in accordance to many navigation systems, where inertial measurements are dead-reckoned in similar ways. The alternative is to extend the state vector with speed and angular velocity, but this increased state dimension would make the particle filter less efficient unless some Rao-Blackwellization is used.

3.2 A Complete Matlab Algorithm

Suppose that an input sequence $u_{1:N}$ or speed and angular velocity (yaw rate), and a likelihood map similar to the one in [Fig. 16.2d](#) are given. The likelihood function $L(i, j)$ is assumed to be represented with a matrix where each row i corresponds to the corresponding element $X(j)$ in the vector X , and similarly for the column $Y(j)$ for a vector Y . A likelihood function such as the one in [Fig. 16.2](#) can be generated from an arbitrary map using the code in Listing 1.

Listing 1: Matlab code for generating likelihood from a bitmapped map

```
y = imread('valla.png'); % Snapshot of map
ys=sum(y,3); % r+g+b
ind=find(ys~=761); % White rgb value in map
yr=zeros(size(ys)); % 0 for non street areas
yr(ind)=1; % 1 for street areas
ym=locmin2(yr,4); % Special: local min over 9x9 square
L=conv2(1-ym,ones(20,20)); % LP-smoothing gives likelihood
surf(L);
shading interp;
campos=[-1000 -4000 10000];
```

The function `locmin2` is nonstandard but simple. It computes the local minimum over a square of nine times nine pixels (within a distance of four pixels). Using this likelihood and some additional parameters for the coordinate transformation from pixels to world coordinates, the particle filter can be implemented in Matlab as given in Listing 2.

Listing 2: Complete Matlab listing for positioning.

```
function Xhat = MapAidedPositioning(y,u,L,pe,vrand,f,h,p0,dp)
Tf      = size(u,2);           % Number of data
N       = 1000;               % Particles
[IndX,IndY] = find(L>0.5*max(L(:))); % Thresholding
IndR    = ceil(length(IndX)*rand(N,1)); % Random road points
Psi     = 2*pi*rand(N,1);     % Random heading

% Coordinate transformation
X = [0:size(L,2)-1] * dp(1) + p0(1);
Y = [0:size(L,1)-1] * dp(2) + p0(2);

% Initialization
Xp(1,:) = IndX(IndR) * dp(1) + p0(1);
Xp(2,:) = IndY(IndR) * dp(2) + p0(2);
Xp(3,:) = Psi;

for k = 1:Tf
    % Road likelihood
    w = interp2(X,Y,L,Xp(1,:),Xp(2,:), 'nearest',0);
    w = w.*pe(y(:,k),Xp);
    % Measurement likelihood
    w = w/sum(w); % Normalization
    Xhat(:,k) = w(:) * Xp'; % Mean estimate
    Xp = resample(Xp,w); % Resampling
    vk = vrand(N); % Process noise
    Xp = f(Xp,u(:,k),vk); % State prediction
end
```

The code in Listing 2 is complete, except for the basic resampling function `resample`. Implementations and a discussion on this function are found in (Gustafsson 2010b). Initialization is performed over the whole road network. The PF is the simplest possible bootstrap (SIR) one originally proposed in (Gordon et al. 1993), and there are more advanced ones that can be more efficient for this application, see Gustafsson (2010b) for a more thorough treatment of implementation and code aspects.

Finally, if there are more sensor information relating to position (such as temporary GPS positions) or course (such as a compass), these are easily incorporated in the filter as additional likelihood function multiplications.

3.3 Tracking Model

The simplest possible tracking model, yet one of the most common ones in applications, is given by a two-dimensional version of Newton's force law:

$$x(t) = \begin{pmatrix} X(t) \\ Y(t) \\ \dot{X}(t) \\ \dot{Y}(t) \end{pmatrix}, \quad \dot{x}(t) = \begin{pmatrix} \dot{X}(t) \\ \dot{Y}(t) \\ w^X(t) \\ w^Y(t) \end{pmatrix} \quad (16.14a)$$

The corresponding discrete time model is given by

$$x_{k+1} = \begin{pmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} x_k + \begin{pmatrix} T^2/2 & 0 \\ T & 0 \\ 0 & T^2/2 \\ 0 & T \end{pmatrix} \begin{pmatrix} w_k^X \\ w_k^Y \end{pmatrix}. \quad (16.14b)$$

Suppose the sensor model depends on the position only, similarly to (16.6),

$$y_k = h(X_k, Y_k) + e_k. \quad (16.15)$$

Since the motion model is linear in the state and noise, the MPF applies, so the velocity component can be handled in a numerically very efficient way.

The code in Listing 3 gives a fundamental example that fits GPS measurements to a road map using a constant velocity model.

Listing 3: Example of how to use the function in Listing 2 for tracking using GPS measurements.

```
% Measurement model.
h=inline('sum(( repmat(y, size(x,2)) - [x(1,:); x(2,:)]).^2 ),1');
pe=inline('exp(-0.5*sum(( repmat(y,1, size(x,2)) - x(1:2,:)).^2)/10^2)', 'y', 'x');

% Dynamic model.
f=inline(['[ x(1,:)+u(1,:)+0.5*v(1,:).*cos(x(3,:)); ', ...
          ' x(2,:)+u(2,:)+0.5*v(1,:).*sin(x(3,:)); ', ...
          ' x(3,:)+v(2,:); '], 'x', 'u', 'v');
vrand=inline('randn(2,N)');

% PF
Xhat=MapAidedPositioning(GPS,[0;0],L,pe,vrand,f,h,[0 0],[0.5 0.5]);
```

3.4 Manifold Model

We here return to the filter framework discussed in Sect. 2.1.3. The manifold model is essentially a one-dimensional version of the tracking model (16.14b),

$$x(t) = \begin{pmatrix} l(t) \\ \dot{l}(t) \end{pmatrix}, \quad \dot{x}(t) = \begin{pmatrix} \dot{l}(t) \\ w(t) \end{pmatrix}, \quad (16.16a)$$

with a discrete time counterpart

$$x_{k+1} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} x_k + \begin{pmatrix} T^2/2 \\ T \end{pmatrix} v_k. \quad (16.16b)$$

With a particle representation of the position l_k along the current road segment, the event of passing a junction is easily detected ($\dot{l}_k < 0$ or $\dot{l}_k > L^i$), and the new road segment can be initialized according to the prior probability π_{ji} . Note that the velocity component \dot{l}_k can be represented with a KF according to the MPF algorithm since the velocity is linear and Gaussian in the model. The posterior then resembles the expression in (16.7).

The same sensor model as in (16.6) and (16.15) now also includes a transformation

$$y_k = h(X^i(l_k), Y^i(l_k)) + e_k, \quad (16.17)$$

where $X^i(l)$ is the mapping from the driven distance l on road segment i to the Cartesian position X , and similarly for $Y^i(l)$.

4 Map Handling

This section describes the fundamentals of vectorized road maps, and some key calculations needed in (dynamic) map matching.

4.1 The Shape Format

In a geographic information system (GIS) different forms of geographically referenced information can be analyzed and displayed. There are two classical methods to store GIS data: raster data (images) and vector data. Different geometrical types can be described by vector data, and basically there are three broad type categories; zero-dimensional points are used to represent points-of-interest, lines are used to represent linear features such as roads and topological lines, and polygons are used to represent particular areas such as lakes. There exist many approaches to store geospatial vector data, and one popular representation is the ESRI shape file. There are 14 different shape types, for example, a road network is represented as a number of PolyLines. A PolyLine is an ordered set of vertices that consists of one or more parts. A part is a connected sequence of two or more points. Parts may or may not be connected to one another. Parts may or may not intersect one another. See ESRI (1998) for more details. Apart from the vector information, each data item may also have attributes that describe properties of the item. Examples of attributes in the road case are road type, street name, speed limit, driving direction, etc. Here, only road network information is considered, but of course there are other types

of information such as terrain type and topological data that can facilitate target tracking and sensor fusion.

For target tracking purposes, it is convenient to have a slightly different representation with redundant information to facilitate and speed up the data processing. One data structure represents the roads, and this structure contains the road stretch and the corresponding attributes. This structure is more or less the raw shape data plus an ID number for each road and an intersection ID for each road end. An additional structure is used for the intersections, and it contains the location and all connected roads (IDs) of each intersection. The exact data structure depends on what type of additional information is included, such as driving direction and prior probabilities for roads in an intersection. The described road structure contains the following fields:

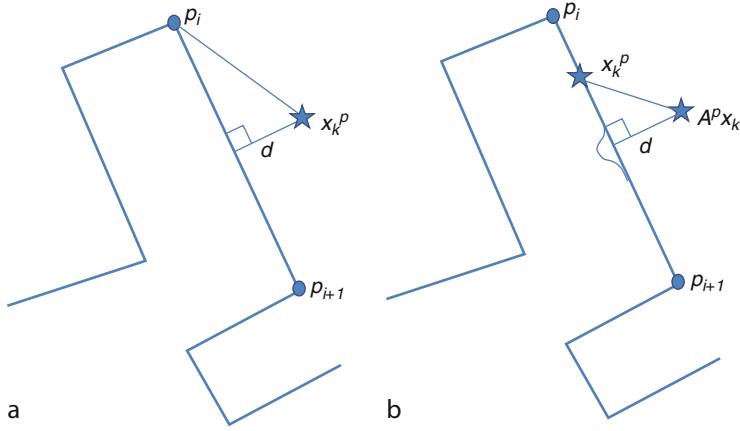
- ID – unique road ID
- N – number of parts
- x – $(1 \times N)$ vector with x coordinates
- y – $(1 \times N)$ vector with y coordinates
- z – $(1 \times N)$ vector with z coordinates
- i_1 – intersection ID of the road-start intersection
- i_2 – intersection ID of the road-end intersection and the intersection structure contains
- ID – unique intersection ID
- M – number of connecting roads
- r – $(1 \times M)$ vector with IDs of the connecting roads
- x – x coordinate
- y – y coordinate
- z – z coordinate

4.2 Computational Issues Related to the Map

Even though the shape format allows for curved road segments, all available maps use the straight line representation. This means that for instance roundabouts are approximated using a number of straight lines. This section will describe a couple of computations needed in road-assisted navigation.

First, consider the virtual measurement approach. If the likelihood is represented as the grid in [Fig. 16.2d](#), then the measurement update is based on interpolation in this grid. This approach requires certain pre-computations and a large memory. A more efficient approach is based on defining a likelihood function based on the distance d to the closest road point. The likelihood can for instance be defined as $l(d) = e^{-\frac{d^2}{2\sigma^2}} + l_0$, where l_0 is optionally added to allow for off-road driving.

Denote the shortest distance to road segment i with d_i , see [Fig. 16.7a](#). Using the scalar product, it is given by



■ Fig. 16.7

Computational issues related to the trajectory in [Fig. 16.1](#). (a) The closest distance d to the road network needs to be computed in the virtual measurement approach. (b) A random noise that takes any prediction back to the road network needs to be generated in the state constraint approach

$$d_i = x_k^p - p_i - \frac{(x_k^p - p_i)^T (p_{i+1} - p_i)}{(p_{i+1} - p_i)^T (p_{i+1} - p_i)} (p_{i+1} - p_i). \quad (16.18)$$

Note that this value must satisfy $d_i \in [0, L_i]$ to be feasible. This calculation has to be performed for all road segments to find the minimum $d = \min_i d_i$. Clearly, an efficient database handling is required. Further, there is a need for an efficient pre-scan of a suitable candidate set of road segments. Here, the absolute norm to each end point can be used, $\|x_k^p - p_i\| = |x_k^x - p_i^x| + |x_k^y - p_i^y|$.

Note that the above operation is also needed in standard map matching using vectorized maps.

The second approach to road-assisted navigation is based on a state constraint in the prediction step, and this can be rather tricky. Consider a linear motion model, for instance the one in [\(16.14b\)](#),

$$x_{k+1} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x_k + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} v_k, \quad (16.19)$$

where the upper blocks A_1 and B_1 correspond to position. Now, if $v_k \in N(0, \sigma^2 I^2)$, we need to generate constrained samples from this distribution that assures that $A_1 x_k + B_1 v_k$ corresponds to a point on the road network. [Figure 16.7b](#) illustrates the geometry for this directional process noise. In this case, with a linear road segment, the conditional distribution can be computed analytically. The result is a one-dimensional Gaussian distribution where the variance is smaller than σ^2 , the larger d the smaller variance.

5 Navigation Applications

This section surveys some approaches to onboard navigation systems based on dead-reckoning. It explains the basic sensor models and provides some illustrative examples from field tests. Such systems can be used as a support or backup to satellite-based navigation.


5.1 Odometric Approach

Odometry is the term used for dead-reckoning the rotational speeds of two wheels on the same axle of a vehicle. It is used in a large range of robotics applications as well as in some vehicle navigation systems. Odometric navigation or positioning based on inertial sensors or dead-reckoning sensors is challenging due to drift and bias in the measurements, particularly for relatively cheap sensors that are available in ordinary passenger vehicles. Map-aided positioning based on vectorized road charts in combination with information from internal automotive sensors such as individual wheel speed and yaw rate available from the CAN-bus has been studied in several Master's theses (Hall 2001; Hedlund 2008; Kronander 2004; Svenzén 2003) and made commercially available at Nira Dynamics AB.

In Hall (2001), the basic theory and implementation for map-aided vehicle positioning were studied, where the particle filter was demonstrated to yield sufficiently good navigation performance when incorporating information from yaw rate and wheel speed sensors. The unknown wheel radius parameter estimation problem was also addressed. The raw signals are the angular velocities of the wheels which can be measured by the ABS sensors in cars. The angular velocities can be converted to virtual measurements of the absolute longitudinal velocity and yaw rate (see Gustafsson [2010a] or Chaps. 13 and 14 in Gustafsson [2010c] for details), assuming a front-wheel-driven vehicle with slip-free motion of the rear wheels, as

$$\vartheta^m = \frac{\omega_3 r_3 + \omega_4 r_4}{2} \approx \vartheta + w_\vartheta, \quad (16.20a)$$


$$\dot{\psi}^m = \vartheta^m \frac{2 \frac{\omega_3 r_3}{\omega_4 r_4} - 1}{\frac{\omega_3 r_3}{\omega_4 r_4} + 1} \approx \dot{\psi} + w_{\dot{\psi}}. \quad (16.20b)$$

See  Fig. 16.8 for the notation. The noise terms can be assumed to be Gaussian,

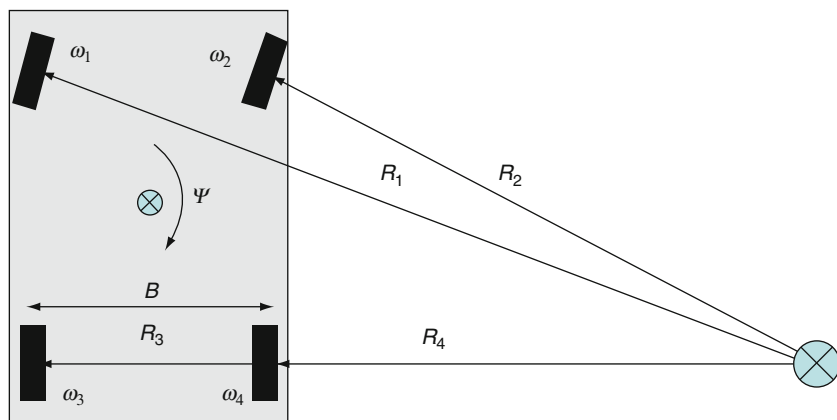
$$w_\vartheta \sim \mathcal{N}(\delta_\vartheta, \sigma_\vartheta^2), \quad (16.21a)$$

$$w_{\dot{\psi}} \sim \mathcal{N}(\delta_{\dot{\psi}}, \sigma_{\dot{\psi}}^2). \quad (16.21b)$$

We assume in this section that both the mean and the variance are known.

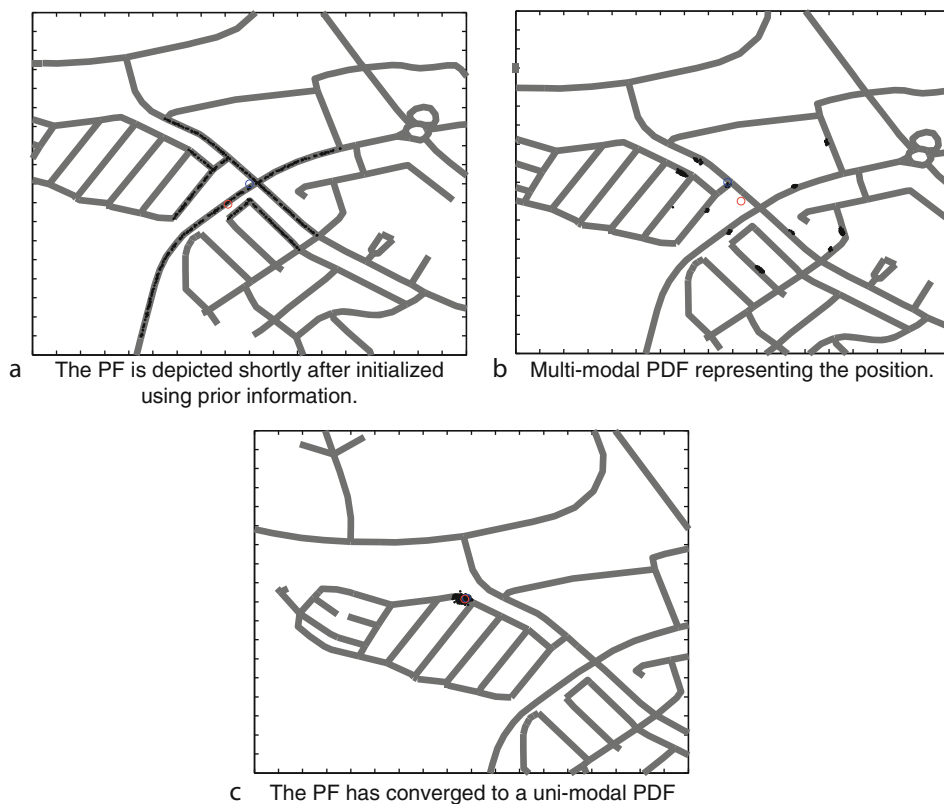
 Figure 16.9 shows an example.

During this development stage, different hardware platforms were investigated. In Svenzén (2003), various particle filter variants were studied, and the MPF/RBPF was used



■ Fig. 16.8





Notation for the lateral dynamics and curve radius relations for a four-wheeled vehicle




■ Fig. 16.9

Map-aided positioning using wheel speed sensor information in combination with road-map information

as an efficient method to incorporate more states in real-time. The ability to enhance positioning while driving slightly off-road was demonstrated in Kronander (2004). This is important for handling inaccurate maps or when the map information is not available. In Hedlund (2008) the positioning aspect was shifted from the use of internal data such as wheel speed information to the case when inertial measurements are available from an external IMU sensor. Typically, the problem studied used accelerometers and gyros as a stand-alone sensor for positioning in combination with the road map. Several different models were studied also in combination with the available internal automotive sensors.

In  Fig. 16.9, the map-aided positioning using wheel speed information and road-map information is demonstrated, where GPS information is used as a ground truth reference only. In  Fig. 16.9a, the particle filter is initialized in the vicinity of the GPS position (blue circle) at a traffic light. The initial distribution is uniform on road segments in a region around the GPS fix. As seen, the GPS position is located slightly off-road, which could be due to a measurement error, multipath phenomenon, or that the road segment width does not match the actual road width. To ensure a robust algorithm against these issues, particles are allowed slightly off-road. The expected mean from the particle filter (red circle) is far away from the true position. In  Fig. 16.9b, the algorithm has been active for some time, and the vehicle has turned left at the crossing. As seen, the PDF is now highly multimodal. The PF algorithm uses only wheel speeds from the CAN-bus. The GPS is only used to evaluate the ground truth. In  Fig. 16.9c, after some turns, the filter has converged to a unimodal distribution and the mean estimate is close to the GPS position.


5.2 Odometric Approach with Parameter Adaptation

The offsets in ( 16.21a, b) depend on the wheel radii as

$$\delta_{\theta} \propto r_3 + r_4, \quad (16.22a)$$

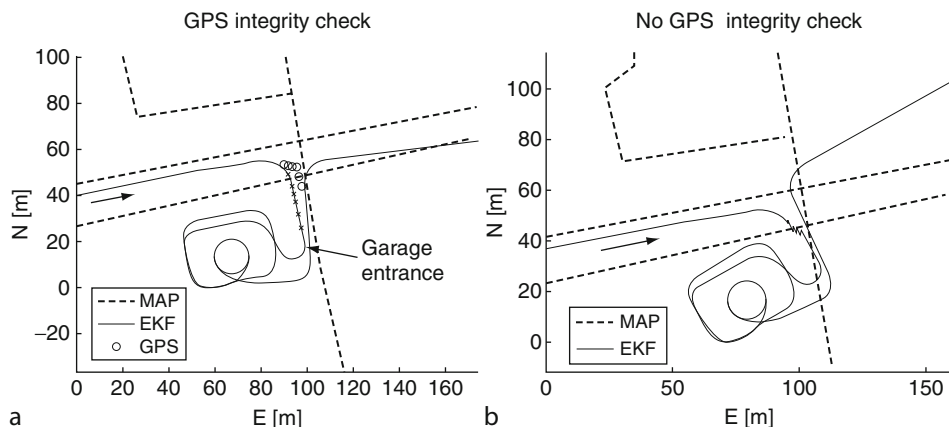
$$\delta_{\psi} \propto r_3 - r_4. \quad (16.22b)$$

Further, the noise variances depend on the surface. Both surface and wheel radii change over time, but with different rates, and the offsets are crucial for dead-reckoning performance.

The standard approach to deal with this problem is to augment the state vector with these two offsets (or the more physical wheel radius offsets). These parameters are then estimated adaptively in the filter.  Figure 16.10 illustrates one advantage with this approach: dead-reckoning improves to enable accurate GPS integrity monitoring so that small GPS errors are detected.

5.3 Inertial Measurement Support

The drawback with the approaches in the preceding sections is that they require wheel speed signals. This is not easy for portable and after-market solutions. An appealing approach is to base the dead-reckoning on an inertial measurement unit (IMU). This can



■ Fig. 16.10

GPS-supported odometry. Multipath propagation close to the parking house gives unreliable GPS positions (marked in (a)). (a) Bias adaptation gives the predictive performance required to exclude GPS outliers. (b) Illustration of what happens if GPS outliers are used in the filter

either be three-dimensional (two horizontally mounted accelerometers and a yaw-rate gyro) or a full six-dimensional unit with three accelerometers and three gyros. The former assumes a flat world and no roll and pitch dynamics of the vehicle, while the latter allows for a more flexible and versatile full-state estimation framework. Further, just one single lateral accelerometer can be used to detect cornering, an important event in dynamic map matching. We here summarize the results in Hedlund (2008).

► [Figures 16.11](#) and ► [16.12](#) from Hedlund (2008) illustrate some different combinations, providing the following conclusions:

- The wheel speed-based dead-reckoning gives superior performance.
- Pure dead-reckoning of an IMU cannot be used as a backup solution to GPS, unless the initial alignment and offset estimation are improved, see ► [Fig. 16.11b](#).
- Dynamic map matching based on dead-reckoning of an IMU and cornering detection from lateral accelerometer is a feasible backup solution to GPS, see ► [Fig. 16.11b](#). However, the robustness is not convincing, see the first plot in ► [Fig. 16.12](#).
- When wheel speed signals are available, also including IMU does not improve the result of dynamic map matching significantly, see ► [Fig. 16.11c, d](#) and the last two plots in ► [Fig. 16.12](#). However, for some driving conditions, this might be the case.

In Hedlund (2008), several different models were studied to find a feasible real-time implementation with sufficient flexibility and performance. The studied models are summarized below:

- Model M0: A pure odometric model (see ► [Sect. 3](#))
- Model M1: A general 3D constant acceleration model with quaternions for orientation representation and accelerometer and rate gyro biases (22 states)

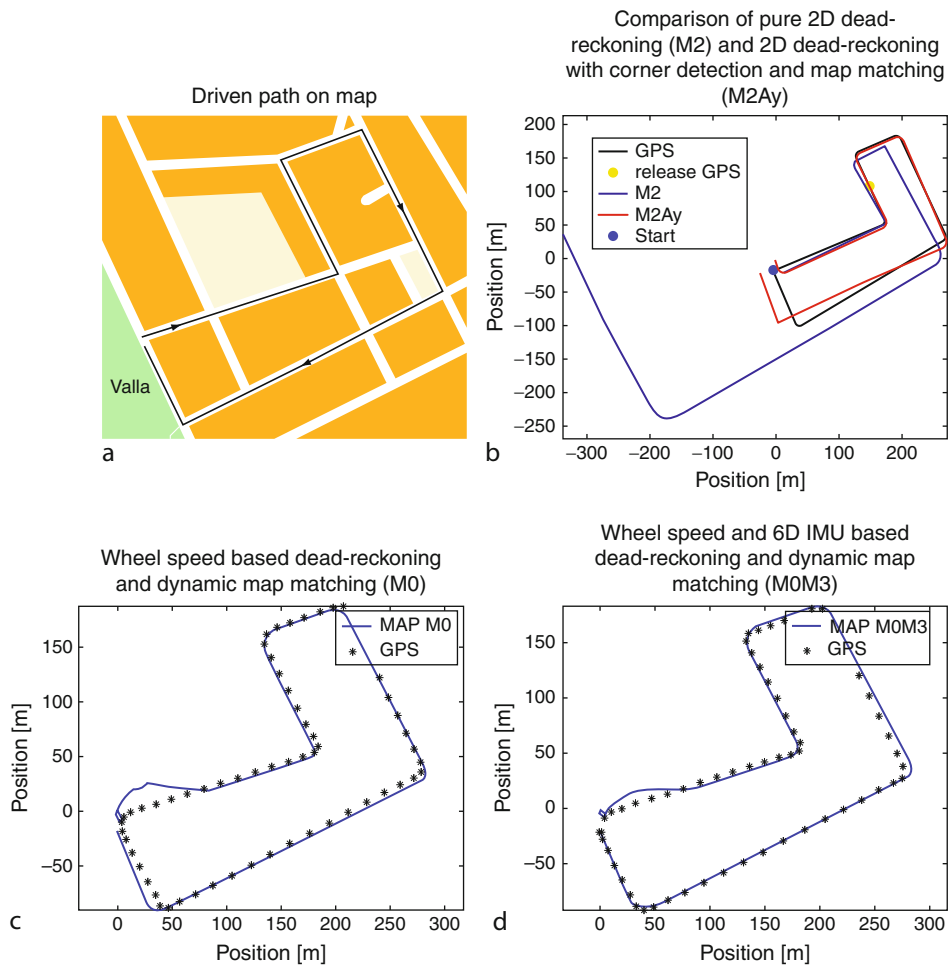


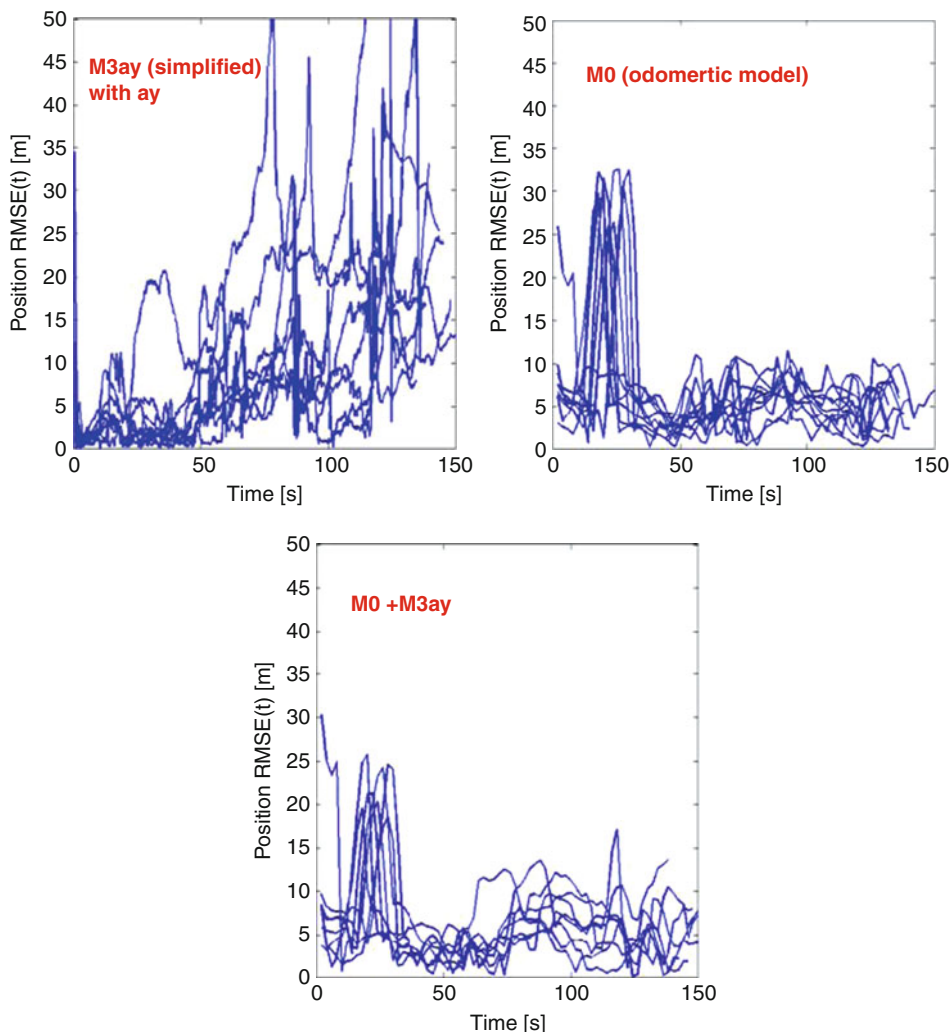
Fig. 16.11

IMU-supported navigation (Pictures from Hedlund 2008)

- Model M2: Utilized the 2D property of vehicle positioning, and it used a coordinated-turn model (11 states)
- Model M2ay: The same as M2 but also using the lateral accelerometer to improve positioning
- Model M3: A simplified coordinated-turn model without biases (six states)

In Model M1, the following state vector is used

$$x = (p \ v \ a \ q \ w \ a_{\text{bias}} \ w_{\text{bias}})^T, \quad (16.23)$$



■ Fig. 16.12

The test drive in [Fig. 16.11a](#) is repeated many times, so the average performance and robustness of the alternatives can be compared (Picture from Hedlund 2008)

where p is the 3D position vector, v the 3D velocity vector, a the 3D acceleration vector, q the four components of the quaternion vector representing orientation, and ω is the angular rate. The continuous time dynamics for the model is given by

$$\dot{x}(t) = (v \ a \ w_a - 0.5S(w)q \ w_w \ w_{a,bias} \ w_{w,bias})^T, \quad (16.24)$$

where w_a is the noise for the acceleration, w_w is the noise for the rotation, and $S(w)$ a rotation matrix.

In order to simplify the dynamics and adapt the model to a more common 2D vehicle scenario, Model M1 is simplified. Basically, a coordinated-turn model is used, where only the longitudinal speed and acceleration are considered as states together with 2D position and a full orientation description with bias terms. This leads to 11 states in Model M2 which describes the vehicle positioning problem very well. The model can be improved by supporting it with information from the lateral accelerometer a_y (Model M2ay). If further reduction of the state vector is needed, the bias terms can be considered as fixed parameters, which will only be updated at the initial alignment. Furthermore, IMU signals can be considered as input signals, yielding only six states. For details we refer to Hedlund (2008).

► [Figure 16.11](#) shows some comparative results from field trials in the same area as in the map of ► [Fig. 16.2](#). ► [Figure 16.12](#) compares the performance over ten different experiments on the same route. As seen the pure IMU model has worse performance than the odometric model. This is because it was hard to estimate the biases and perform a sufficiently good initial alignment, compensating for the gravitational vector. This can probably be improved; however, since noisy signals are double integrated to yield position, it is a much tougher problem than integrating wheel speed signals once. As seen a sensor fusion between automotive sensors and IMU yields basically the same result as the odometric one.

From ► [Figs. 16.11](#) and ► [16.12](#), we conclude the following:

- The wheel speed-based dead-reckoning gives superior performance.
- Pure dead-reckoning of IMU cannot be used as a backup solution to GPS, unless the initial alignment and offset estimation is improved, see ► [Fig. 16.11b](#).
- Dynamic map matching based on dead-reckoning of IMU and cornering detection from lateral accelerometer form a feasible backup solution to GPS, see ► [Fig. 16.11b](#). However, the robustness is not convincing, see the first plot in ► [Fig. 16.12](#).
- When wheel speed signals are available, also including IMU does not improve the result of dynamic map matching significantly, see ► [Fig. 16.11c, d](#) and the last two plots in ► [Fig. 16.12](#). However, for some driving conditions, this might be the case.

6 Tracking Approaches

Tracking road-bound vehicles is important in surveillance and certain applications in intelligent transportation systems. This section illustrates how this can be done using wireless radio measurements, microphone sensor networks, radars and airborne video cameras.

6.1 Radar Support

A radar system emits electromagnetic waves and analyzes the reflected waves to determine the range, direction, and radial speed of both moving and fixed objects. Both the initial Kalman filter-based studies (Kirubarajan et al. 2000; Shea et al. 2000a, b) and subsequent

particle filter-based approaches (Arulampalam et al. 2002; Ristic et al. 2004) on target tracking with road network information were motivated by radar applications. In this direction of research, extensive effort was spent for improving the methods (Cheng and Singh 2007; Koller and Ulmke 2007; Payne and Marrs 2004; David Salmond et al. 2007; Skoglar et al. 2009; Ulmke and Koch 2006), especially for ground moving target indication (GMTI) which is a mode of operation of a radar system where the range rate (Doppler) is used to discriminate moving targets against stationary clutter. The probability of detection in GMTI systems depends not only on the environment topography, but also on the relative radial velocity of the target.


Let $x_k = (X_k, Y_k)^T$ be the position of the target relative a global Cartesian reference system. For simplicity, assume that the sensor is located at the origin. The GMTI observation model can be expressed as

$$y_k = h(x_k, u_k, e_k) = \begin{pmatrix} r_k \\ \theta_k \\ \dot{r}_k \end{pmatrix} + e_k = \begin{pmatrix} \sqrt{X_k^2 + Y_k^2} \\ \arctan_2(Y_k, X_k) \\ \frac{X_k \dot{X}_k + Y_k \dot{Y}_k}{\sqrt{X_k^2 + Y_k^2}} \end{pmatrix} + e_k \quad (16.25)$$

where e_k is the measurement noise modeled as

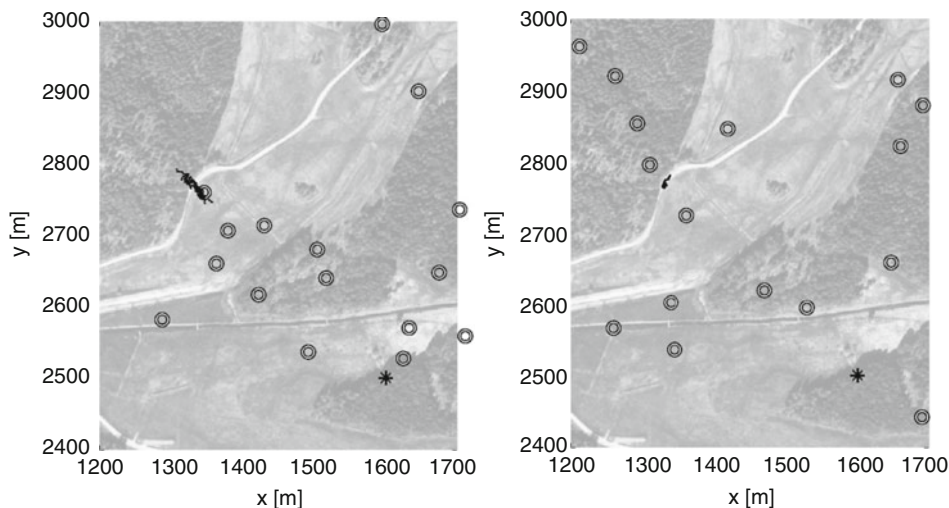
$$e_k \sim \mathcal{N}(0, \text{diag}(\sigma_r^2, \sigma_\theta^2, \sigma_{\dot{r}}^2)). \quad (16.26)$$

This is just a 2D model, but it is straightforward to include the elevation angle to get a 3D description.

A basic simulation example is here given to show the advantages of using road network information when tracking a moving on-road target. The target is detected if the radial speed is above the minimum detectable velocity (MDV). False detections are assumed to be uniformly and independently distributed and the number of false detections is assumed to be Poisson distributed. A global nearest neighbor (GNN) association algorithm is used with standard gating and initiator logic. Snapshots from two GMTI road target tracking examples are shown in  Fig. 16.13, where an off-road target model (left) and an on-road target model (right) are used, respectively. A particle filter is used in both cases, and it is possible to see that an on-road model is advantageous since the resulting particle cloud is significantly denser, i.e., the variance is smaller. The advantage of using road information is even more obvious when the target is not detected, e.g., due to Doppler blindness, and the filter has to predict the target motion. More extensive analysis of the target tracking problem with GMTI can be found in Arulampalam et al. (2002), Ulmke and Koch (2006), and references therein.

6.2 Wireless Radio Network Support

There are a number of different measurement types that can be used to position a wireless network user. Most of the work focuses on the range measurements depending on time of



■ Fig. 16.13

Snapshots from two GMTI road target tracking examples with an off-road target model (*left*) and an on-road target model (*right*). The stationary radar sensor is located near the lower right corner and the circles indicate all detections, both false and true. A particle filter is used, and the particles are represented by dots. An on-road model is advantageous since the resulting particle cloud is significantly denser, i.e., the variance is smaller. In particular, when the target is not detected due to Doppler blindness, the prediction of the target motion is better when using the road information

arrival (TOA), time difference of arrival (TDOA) observations, and received signal strength (RSS) observations, see Gustafsson and Gunnarsson (2005) and the references therein. Among such alternatives, the RSS measurements, which do not require any additional hardware, are the most easily available. The RSS measurement might, on the other hand, be much more noisy than other type of measurements and prior information like the road maps proves to be crucial in obtaining a reasonable performance. The two most common models connecting the target range to the RSS measurements are the general exponential path loss model, which is known as Okumura-Hata model (Hata 1980; Okumura et al. 1968), and a dedicated power map constructed off-line for the region of interest.


The Okumura-Hata model says that the RSS value in a log power scale decreases linearly with the log-distance to the antenna, i.e.,



$$z_k = P_{BS} - 10\alpha \log_{10} \left(\|p_{BS} - x_k^p\|_2 \right) + e_k, \quad (16.27)$$



where z_k is the RSS measurement; x_k^p is the target position; P_{BS} is transmitted signal power (in dB); α is the path loss exponent; e_k is the measurement noise; p_{BS} is the position of the antenna (base-station (BS)); and the notation $\|\cdot\|_2$ denotes the standard ℓ_2 -norm. This is quite a crude approximation, where the noise level is high and further depends on multipath and non-line-of-sight (NLOS) conditions.

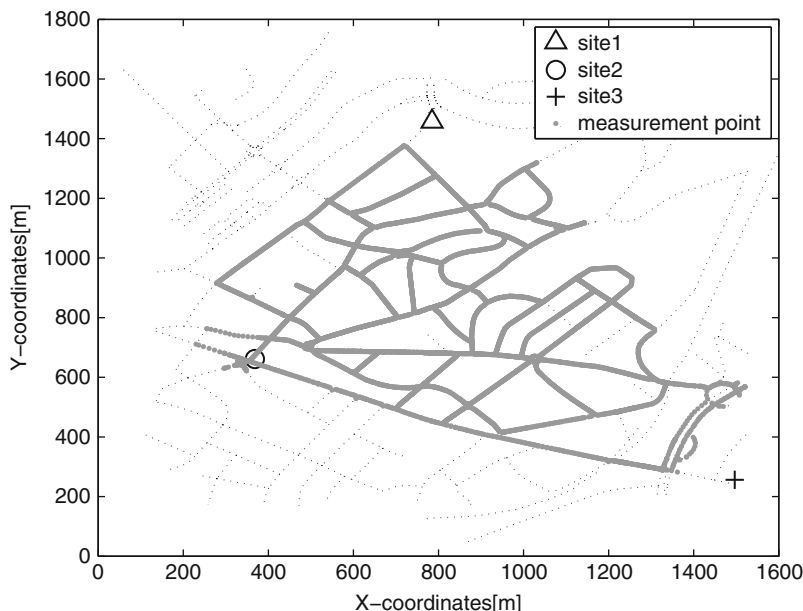
The second alternative is to determine the RSS values at discrete points in the area of surveillance and save this in a database. This can be done using off-line measurement campaigns, adaptively by contribution from users or using cell planning tools. The advantage of this effort is a large gain in signal to noise ratio and less sensitivity to multipath and NLOS conditions. The set of RSS values that are collected for each position from various BSs is called the fingerprint for that location. The idea of matching observations of RSS to the map of the previously measured RSS values is known as fingerprinting. The set of all fingerprints forms an unconventional but informative measurement model, and this can be used in localization.

With both alternatives, there are two possible approaches: static and dynamic localization. In the static approach, no assumptions about the target motion are made, and for each measurement, one generates a position estimate based on the corresponding measurement. In this case, the estimator is a static function of the input measurement. In the second approach, one can use a motion model for the target behavior along with an estimator (which is a dynamic function of the measurements) sequentially updating the estimates with the incoming measurements. Since the information in different measurements is fused by means of the motion model with such an approach, much better accuracy is achievable.

In this section, we present the results obtained in the example study presented in (Bshara et al. 2010), where WiMAX RSS data from three sites and seven BSs was collected in the urban area (in Brussels) shown in  Fig. 16.14. The collected data was saved in a database and used to locate the target based on a separate test data. The following five approaches were used to localize the target.

- Static Positioning: The test data is used in a static manner as described above to locate the target. Basically, the position of the closest fingerprint to the collected measurement becomes the estimate.
- Dynamic–OH–off-road: A particle filtering–based approach with an off-road (i.e., no road-map information) target motion model and the OH-model of ( 16.27) as the measurement model.
- Dynamic–OH–on-road: A particle filtering–based approach with a road-constrained target motion model and the OH-model of ( 16.27) as the measurement model.
- Dynamic–Fingerprinting–off-road: A particle filtering–based approach with an off-road (i.e., no road-map information) target motion model and the fingerprint database as the measurement model. See the details on how to utilize the fingerprints in the measurement likelihood calculation in (Bshara et al. 2010).
- Dynamic–Fingerprinting–on-road: A particle filtering–based approach with a road-constrained target motion model and the fingerprint database as the measurement model. See the details on how to utilize the fingerprints in the measurement likelihood calculation in (Bshara et al. 2010).

The results in the form of the cumulative distribution functions of the position estimation errors are shown in  Fig. 16.15 in two plots. The dynamic and static estimation results are compared in  Fig. 16.15a. A definite advantage of the dynamic approaches is seen even without the road-map information with 95% probability, though,



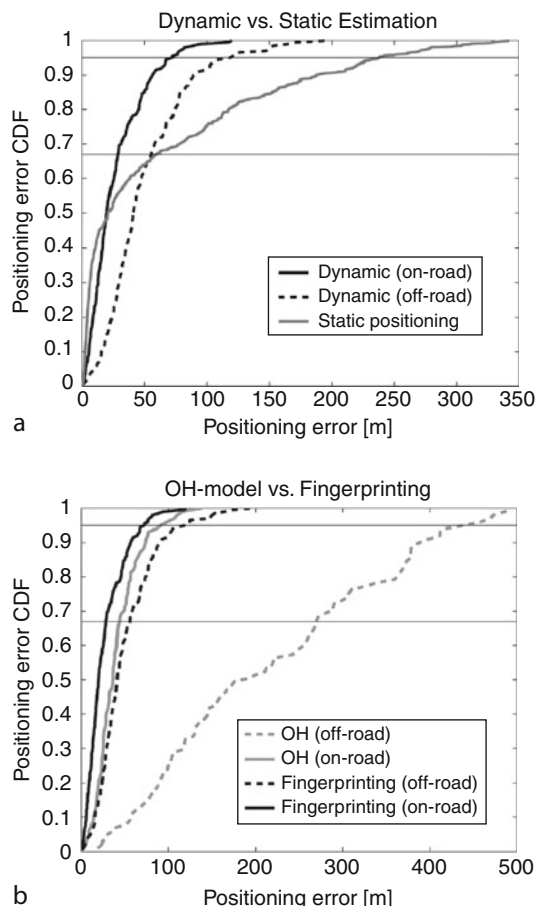
■ Fig. 16.14

MAP of the area under study, sites with base stations, and the measurement locations

in rare events (below 0.67%), static estimation can sometimes get better results. On average, the road information in the dynamic case seems to result in about 25 m better accuracy. ● [Figure 16.15b](#) compares the OH-based and the fingerprinting based methods. Without the road constraints, the simple OH-model behaves much worse than the fingerprinting based methods with positioning errors above 400 m (95% line). However, when the on-road model is utilized, the accuracy can reach up to 100 m (95% line) which is even better than the off-road fingerprinting case. Hence, the road-map information is capable of making even the crudest measurement models behave as good as sophisticated ones.

6.3 Sensor Network Support

This section considers the problem of localizing an unknown number of targets around an acoustic sensor network. Since acoustic power is additive at each sensor, the RSS from different sources cannot be resolved and the framework is difficult to extend to multiple target tracking. In practice, the exponential signal decay rate implies that the closest target will dominate each sensor observation. One applicable approach is to consider each sensor as a binary proximity sensor as studied in for instance (Boers et al. 2008). However, this requires an excessive amount of sensors to get accurate multi-target tracking (MTT). Another problem associated with the RSS measurements is that the emitted acoustic powers from the targets are unknown and must be estimated along with the target states. For these reasons, MTT ideas with acoustic sensors appeared in the literature with either



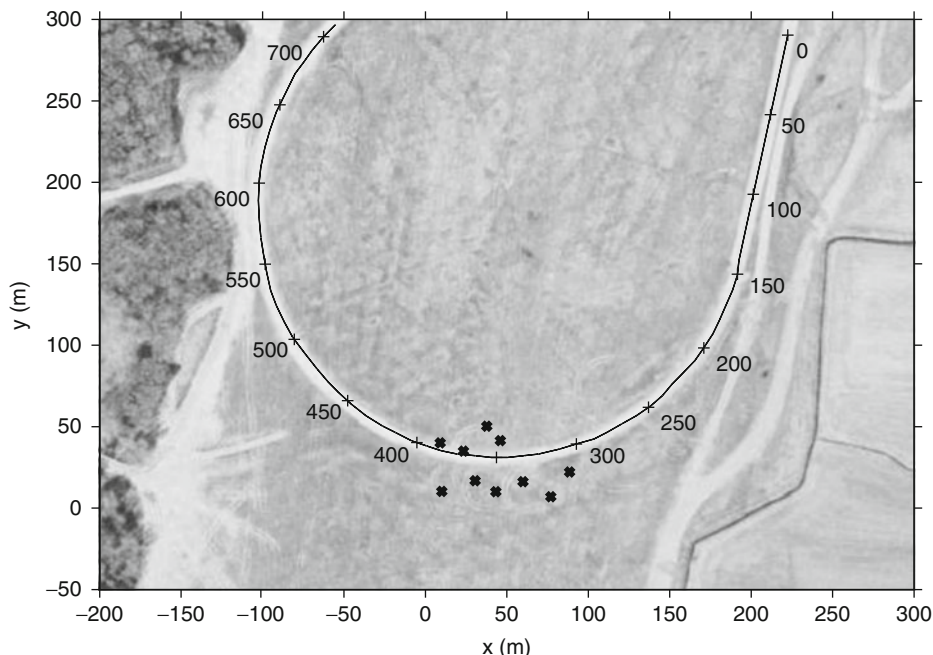
■ Fig. 16.15

Wireless network-supported tracking

direction of arrival (DOA) (Cevher et al. 2007a, b; Fallon and Godsill 2007, 2008) or time difference of arrival (TDOA) (Wing-Kin et al. 2006) measurements. The power (and/or energy)-based measurements case was also examined with few examples in (Bugallo and Djuric 2006; Sheng and Hu 2005) which assume that either the number of targets or the emitted powers are known.

When the road information is supplied, the problem can be tackled much more easily. We here summarize results in Orguner and Gustafsson (2010). The map of the area under study is shown in Fig. 16.16 along with the road information and microphone positions. The on-road position coordinates p_{η} are marked in the figure at each 50 m. We have ten microphones collecting data at 4 kHz placed around the road. Each microphone position is illustrated with a cross sign in Fig. 16.16.

The synchronized recordings of a motorcycle and a car are used while the correct positions are measured with GPS sensors. The correct positions of the targets projected



■ Fig. 16.16

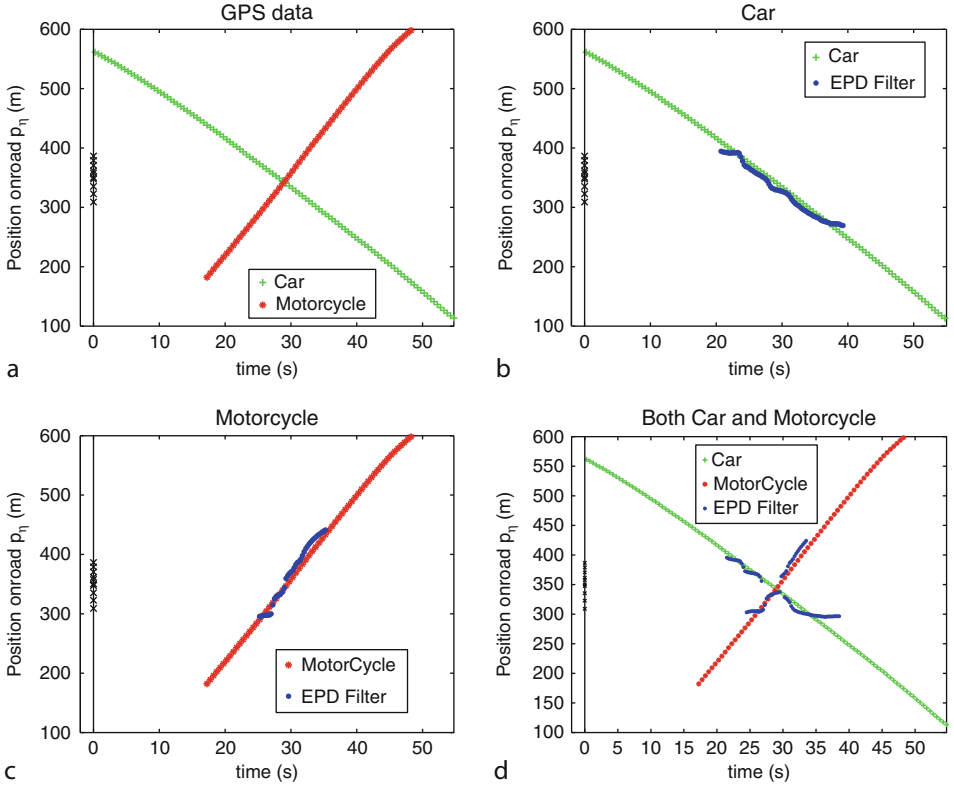
The map of the area, road segment, microphones, and coordinates used in the example. The distance markings on the road segment denote the on-road position coordinates.

Microphone positions are illustrated with cross signs

onto the road coordinates are shown in ► Fig. 16.17a. The microphone network is also illustrated in ► Fig. 16.17a with cross signs at $t = 0$ denoting the closest on-road point to each microphone. The recordings for the motorcycle and the car were obtained separately, and we obtain our two-target data by adding the sound waveforms for the two cases. The on-road position coordinates are discretized uniformly with 5 m distance between adjacent points and the point mass filter-based Emitted Power Density (EPD) filter of Orguner and Gustafsson (2010) is run

1. For car's sound data only
2. For motorcycle's sound data only
3. For the superposed sound data of the car and the motorcycle

The resulting position estimates obtained are illustrated in ► Fig. 16.17b–d, respectively. Single target detection and tracking seem to be good except for some occasional missing detections in the motorcycle only case. In the two-target case, the target initiation delays a little and target loss happens a little earlier. However, both targets can be tracked quite similarly to the single target cases, which would be very difficult without the road network information if not impossible.



■ Fig. 16.17

The correct on-road positions of the two targets and the estimation results. The closest on-road point to each microphone is also illustrated with cross signs at $t = 0$

6.4 Vision Support

Vision sensors are bearings-only sensors providing the azimuth and inclination to the target relative the sensor platform. A vision sensor is here defined as a staring-array electrooptical/infrared sensor (EO/IR) with limited field-of-view (FOV). Let $p = (X, Y, Z)^T$ be the 3D position of the target relative a global Cartesian reference system. For simplicity, assume that the sensor is located at the origin. An observation at time t is the relative angles between the sensor and the target, i.e.,

$$y_k = h(x_k, u_k, e_k) = \begin{pmatrix} \phi_k \\ \theta_k \end{pmatrix} + e_k = \begin{pmatrix} \arctan_2(Y_k, X_k) \\ \arctan_2(Z_k, \sqrt{X_k^2 + Y_k^2}) \end{pmatrix} + e_k, \quad (16.28)$$

where e_k is the measurement noise modeled as

$$e_k \sim \mathcal{N}(0, \sigma^2 I_{2 \times 2}). \quad (16.29)$$

A measurement y_k is obtained by transforming a detection at an image point $(u \ v)^T$ to azimuth and inclination angles given the knowledge of the sensor orientation. For an ideal vision sensor, a point $p^b = (X^b \ Y^b \ Z^b)^T$, expressed in Cartesian coordinates relative to the camera-fixed reference system, is projected in a virtual image plane onto the image point $(u \ v)^T$ according to the ideal perspective projection formula

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{f}{Z^b} \begin{pmatrix} X^b \\ Y^b \end{pmatrix} \quad (16.30)$$

where f is the focal length of the camera. However, in practice, the intrinsic parameters of the vision sensor model must be estimated to handle lens distortion etc.

Prior information about the environment, like the road network, will improve the target-tracking performance significantly. In particular, in the vision sensor case, the problem is not fully observable if the sensor is stationary. However, as the example below indicates, there is more prior information apart from the road network that can support the estimation process. In this simulation example, a single car is tracked by a UAV equipped with a camera. The simulation environment and the path of the car are shown in [Fig. 16.18](#). The sensor platform is flying in a circle with a radius of 100 m and approximately 100 m above the ground, the approximate sensor view is shown [Fig. 16.19](#). An observation is the azimuth and inclination angles obtained from a detection.

The results from three different target tracking filters are shown in [Figs. 16.19](#) and [16.20](#). The first filter is a standard bootstrap PF that assumes that the car will always be on the known road network manifold (this filter is called “on-road PF”). The second filter is a bootstrap PF based on a coordinated-turn-like model that is not using the road network information (this filter is called “off-road PF”). The third filter is a multiple model PF (MMPF) with two sub-filters, one sub-filter identical to the on-road PF and one sub-filter identical to the off-road PF (this filter is called “on/off-road MMPF”). All filters



■ Fig. 16.18

Left: The simulation environment (Rydell et al. 2010). **Right:** The path of the car, driving from left to right. The sensor platform is flying north of this area

have 1,000 particles in total. The root mean square position errors for 100 simulation runs are shown in Fig. 16.20. The car is occluded behind a building between 12 and 17 [s] and the errors grow due to that, especially the off-road PF have serious problem here. The car is rediscovered, but after about 18 [s], the car enters a parking lot that is not part of the road network model. Hence, the on-road PF diverge, but the MMPF and off-road PF can handle that mode change. In Fig. 16.20, the on-road mode probability of the MMPF is shown. When the car is on the parking lot, the on-road probability is very small.

Knowledge about the buildings and vegetation is also very useful to be able to draw conclusions from non-detection (Skoglar et al. 2009). In the current example, the car moves through an intersection just before it is occluded by a building, see Fig. 16.19.

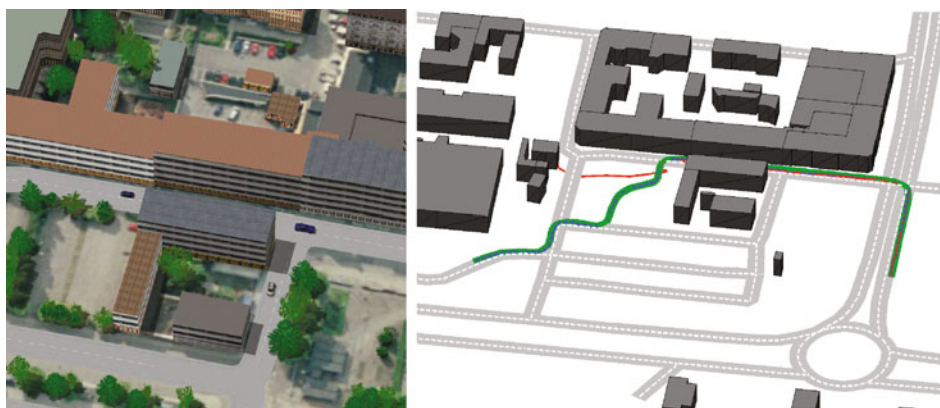


Fig. 16.19

Left: The car to the right of the building will soon be occluded. *Right:* The filter results. The MMPF (dark gray) and the off-road PF (black) are hard to discriminate from the ground truth (gray). The PF (light gray) is diverging when the car is off-road

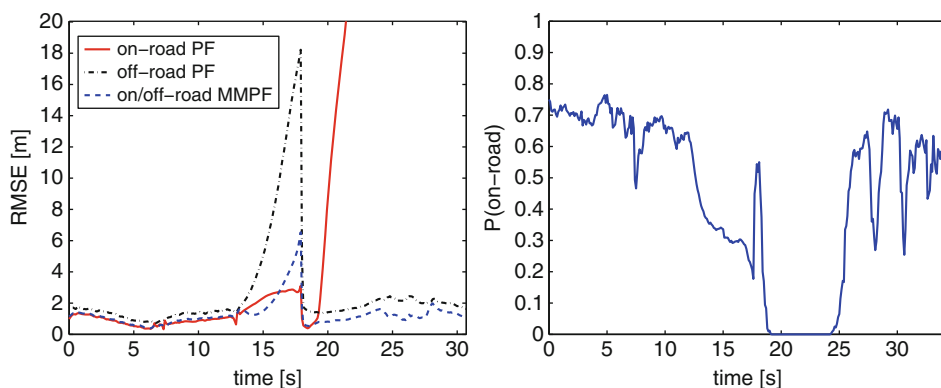


Fig. 16.20

Left: The RMSE results. *Right:* on-road mode probability in the MMPF

A target-tracking filter that is not using this so-called negative information will spread its particles on both roads since no detections are received. However, a filter that utilizes the negative information would discard the particles on the visible road segment, since if the car would have been on that road segment, it should have been detected.

Even though this example is a simulation where some problems, such as navigation error and multiple targets are neglected, it is still possible to draw some general conclusions. Using road network data as prior information will improve the target-tracking performance for sensor vision applications. Especially when the motion of the sensor platform is rather limited since bearings-only tracking performance depends very much on the movement of the sensor platform. The road information is also very useful to predict the target motion in the case of non-detection, for instance due to occlusion. However, algorithms that rely much on prior information should always be used with a fail-safe algorithm that can take over when the prior information is wrong. If the navigation error is slowly varying the measurements will be biased and may cause major problems for the filter that uses the road network as state space manifold, especially when the targets are close to intersections. Using “negative information” is a conceptually simple thing to do to increase the performance in environments and situations where the probability of detection varies. The gain of negative information is more obvious when using it in the road network context.

7 Conclusions

Map matching is an appealing approach to road-assisted navigation when accurate position information is available, for instance from GNSS (global navigation satellite systems). Without a reliable position sensor, the navigation problem is more challenging, and a kind of dynamic map matching is needed that takes both a motion model of the vehicle and the topology of the map into account. We have surveyed state-of-the-art algorithms for road-assisted navigation and tracking using the following main principles for how to incorporate the road-constraint into a filtering framework:

- The road constraint is included as a virtual measurement. This fits the particle filter algorithm well, where the measurement update corresponds to multiplying each weight with a scalar that depends on the distance to the closest road point. The advantage is its simplicity. The disadvantage is the potentially poor particle efficiency, where a large number of positions end up outside the road network.
- The road constraint is converted into a direction process noise that projects the state back to the road network. This is a commonly used approach in tracking applications such as GMTI (ground moving target indicator). The advantages are that it fits a Kalman filter framework and that it provides better particle efficiency in a particle filter than the virtual measurement approach. The disadvantage is that the mathematical operation to generate such noise is quite complex and that ad hoc approximations may be needed.

- The road network is interpreted as a manifold, where a discrete state is used to represent the road network between junctions, and a continuous state variable represents the one-dimensional position between the junctions. The advantage is that this approach utilizes all information in an efficient way. The disadvantage is a more complex algorithm.

A wide range of sensor combination and performance indicators were presented.

The position estimate from dynamic map matching can never be more accurate than the road map itself, and commercial maps are always subject to small deviations from reality. For navigation purposes, this does not pose any problems. For advanced driver assistance systems (ADAS), the position relative to the road and both stationary and dynamic obstacles are needed, and this is the subject of ➤ Chap. 15.

Acknowledgments

This work has been supported by the Swedish Research Council (VR) under the Linnaeus Center CADICS, the VR project grant Extended Target Tracking, the SSF excellence center MOVIII, the Vinnova excellence center FOCUS and the strategic research environments ELLITT in the ICT area and Security Link in the security area.

The authors would also like to thank NIRA Dynamics, SAAB EDS, and FOI for initiating a series of stimulating master thesis in this area. Special thanks to Mussa Bshara for collecting the WiMAX data in ➤ Sect. 6.2.

References

- Arulampalam MS, Gordon N, Orton M, Ristic B (2002) A variable structure multiple model particle filter for GMTI tracking. In: Proceedings of international conference on information fusion, vol 2, Annapolis, MA, pp 927–934
- Blom HAP, Bar-Shalom Y (1988) Interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Trans Autom Control* 33(8):780–783
- Boers Y, Driessen H, Schipper L (2008) Particle filter based sensor selection in binary sensor networks. In: Proceedings of the 11th international conference on information fusion, Cologne
- Bshara M, Orguner U, Gustafsson F, VanBiesen L (2010) Fingerprinting localization in wireless networks based on received signal strength measurements: a case study on WiMAX networks. www.control.lsi.liu.se/fredrik/reports/09vtvmussa.pdf
- Bugallo MF, Djuric PM (2006) Tracking of time-varying number of moving targets in wireless sensor fields by particle filtering. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP), vol 4, Toulouse, pp 865–868
- Cevher V, Sankaranarayanan AC, McClellan JH, Chellappa R (2007a) Target tracking using a joint acoustic video system. *IEEE Trans Multimedia* 9(4):715–727
- Cevher V, Velmurugan R, McClellan JH (2007b) Acoustic multitarget tracking using direction-of-arrival batches. *IEEE Trans Signal Process* 55(6):2810–2825
- Cheng Y, Singh T (2007) Efficient particle filtering for road-constrained target tracking. *IEEE Trans Aerosp Electron Syst* 43(4):1454–1469
- ESRI (1998) ESRI shapefile technical description – an ESRI white paper. <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>
- Fallon M, Godsill S (2007) Multi target acoustic source tracking using track before detect.

- In: Proceedings of IEEE workshop on applications of signal processing to audio and acoustics, New Paltz, NY, pp 102–105
- Fallon M, Godsill S (2008) Multi target acoustic source tracking with an unknown and time varying number of targets. In: Proceedings of the conference on hands-free speech communication and microphone arrays (HSCMA), Trento, pp 77–80
- Gordon NJ, Salmond DJ, Smith AFM (1993) A novel approach to nonlinear/non-Gaussian Bayesian state estimation. In: IEE Proceedings on radar and signal processing, vol 140, pp 107–113
- Gustafsson F, Gunnarsson F (2005) Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements. *IEEE Signal Process Mag* 22:41–53
- Gustafsson F (2010a) Rotational speed sensors: limitations, pre-processing and automotive applications. *IEEE Instrum Meas Mag* 13(2):16–23
- Gustafsson F (2010b) Particle filter theory and practice with positioning applications. *IEEE Trans Aerosp Electron Mag Pt II Tuts* 7(July):53–82
- Gustafsson F (2010c) Statistical sensor fusion. Studentlitteratur, Lund
- Hall P (2001) A Bayesian approach to map-aided vehicle positioning. Master's thesis LiTH-ISY-EX-3102, Department of Electrical Engineering, Linköping University, Linköping
- Hata M (1980) Empirical formula for propagation loss in land mobile radio services. *IEEE Trans Veh Technol* 29(3):317–325
- Hedlund G (2008) Map aided positioning using an inertial measurement unit. Master's thesis LiTH-ISY-EX-4196, Department of Electrical Engineering, Linköping University, Linköping
- Julier SJ, Uhlmann JK, Durrant-Whyte HF (1995) A new approach for filtering nonlinear systems. In: IEEE American control conference, Seattle, Washington, pp 1628–1632
- Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng Trans ASME Series D* 82:35–45
- Kirubarajan T, Bar-Shalom Y, Pattipati KR, Kadar I (2000) Ground target tracking with variable structure IMM estimator. *IEEE Trans Aerosp Electron Syst* 36(1):26–46
- Koller J, Ulmke M (2007) Data fusion for ground moving target indicator. *Aerosp Sci Technol* 11:261–270
- Kramer SC, Sorenson HW (1988) Recursive Bayesian estimation using piece-wise constant approximations. *Automatica* 24:789–801
- Kronander J (2004) Map aided positioning using an inertial measurement unit. Master's thesis LiTH-ISY-EX-3578, Department of Electrical Engineering, Linköping University, Linköping
- Li XR, Bar-Shalom Y (1996) Multiple-model estimation with variable structure. *IEEE Trans Autom Control* 41(4):478–493
- Okumura Y, Ohmori E, Kawano T, Fukuda K (1968) Field strength and its variability in VHF and UHF land-mobile radio service. *Rev Elec Commun Lab* 16(9-10):825–873
- Orguner U, Gustafsson F (2010) Multi target tracking with acoustic power measurements using emitted power density. In: Proceedings of 13th international conference on information fusion (FUSION '10), Edinburgh
- Payne O, Marrs A (2004) An unscented particle filter for GMTI tracking. In: Proceedings of aerospace conference, vol 3, Big Sky, Montana, pp 1869–1875
- Ristic B, Arulampalam S, Gordon N (2004) Beyond the Kalman filter: particle filters for tracking applications. Artech House, London (Chapter 10)
- Rydell J, Haapalahti G, Karlholm J, Näsström F, Skoglar P, Stenborg KG, Ulvklo M (2010) Autonomous functions for UAV surveillance. In: Proceedings of International conference on intelligent unmanned systems (ICIUS), Bali
- Salmond D, Clark M, Vinter R, Godsill S (2007) Ground target modelling, tracking and prediction with road networks. In: Proceedings of International conference on information fusion, Québec
- Schmidt SF (1966) Application of state-space methods to navigation problems. *Adv Control Syst* 3:293–340
- Shea PJ, Zadra T, Klammer D, Frangione E, Brouillard R (2000a) Improved state estimation through use of roads in ground tracking. In: Proceedings of the SPIE conference on signal and data processing of small targets, vol 4048, Orlando, FL, pp 312–332
- Shea PJ, Zadra T, Klammer D, Frangione E, Brouillard R (2000b) Precision tracking of ground targets. In: Proceedings of aerospace conference, vol 3, Big Sky, Montana, pp 473–482

- Sheng X, Hu Y-H (2005) Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE Trans Signal Process* 53(1):44–53
- Skoglar P, Orguner U, Törnqvist D, Gustafsson F (2009) Road target tracking with an approximate Rao-Blackwellized particle filter. In: *Proceedings of 12th International conference on information fusion*, Seattle, Washington
- Sorenson HW, Alspach DL (1972) Recursive Bayesian estimation using Gaussian sums. *IEEE Trans Autom Control* 17:439–448
- Svenzén N (2003) Real time map-aided positioning using a Bayesian approach. Master's thesis LiTH-ISY-EX-3297, Department of Electrical Engineering, Linköping University, Linköping
- Ulmke M, Koch W (2006) Road-map assisted ground moving target tracking. *IEEE Trans Aerosp Electron Syst* 42(4):1264–1274
- Wing-Kin Ma, Ba-Ngu Vo, Singh SS, Baddeley A (2006) Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach. *IEEE Trans Signal Process* 54(9):3291–3304

17 State-of-the-Art In-Car Navigation: An Overview

Isaac Skog · Peter Händel


School of Electrical Engineering, KTH Royal Institute of
Technology, Stockholm, Sweden

1	<i>Principles of In-Car Navigation Systems</i>	437
1.1	Navigation System Building Blocks	437
1.2	Navigation System Figures of Merits	439
2	<i>Global Navigation Satellite Systems</i>	439
2.1	Global Navigation Satellite Systems	439
2.2	GNSS Error Sources	440
2.3	Horizontal Position Accuracy Prediction	441
2.4	Differential GNSS	441
2.5	GNSS Integrity Monitoring	442
2.5.1	Key Steps in Integrity Monitoring	443
3	<i>Vehicle Motion Sensors</i>	443
3.1	Non-Self-Contained Encoder-Based Motion Sensors	444
3.2	Self-Contained Motion Sensors	445
3.3	Errors Sources for Inertial Sensors	446
4	<i>Coordinate Systems</i>	446
4.1	Coordinate Systems	447
5	<i>Information Processing Based on Vehicle Sensors</i>	449
5.1	Dead-Reckoning	449
5.2	Inertial Navigation	449
5.3	Vehicle Models and Motions	452
6	<i>Digital Maps and Map-Matching</i>	453
6.1	Digital Maps	454
6.2	Map-Matching Steps	455

- 7 **Information Fusion** **456**
 - 7.1 Filter Structures 456
 - 7.1.1 Filter Structures for Information Fusion 456
 - 7.2 Filter Algorithms 458
 - 7.3 Information Fusion in Systems Including a GNSS Receiver 459
- 8 **Summary** **460**

Abstract: The basics around in-car navigation is discussed, including the principals of contemporary systems, global navigation satellite system basics, dead-reckoning, map-matching, and strategies for information fusion. In-car navigation system are generally made out of three building blocks, an information source block, an information fusion block, and an user interface block. This chapter presents an overview of the information source block and the information fusion block. First, the ideas of operation and main characteristics of the four most commonly used information sources, global navigation satellite systems, vehicle motion sensors, road maps, and mathematical models of the vehicle dynamics, are reviewed. Thereafter, common techniques to combine the information from the different information sources into an estimate of the position, velocity, etc. of the car are reviewed.


1 Principles of In-Car Navigation Systems

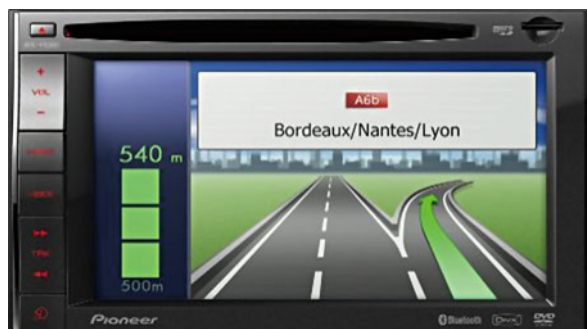
At several markets, there is a high penetration of original equipment manufacturer (OEM) and third-party in-car navigation systems for private users as well as for professional users. A typical consumer head unit is displayed in  Fig. 17.1. The large color screen is the major instruments for machine–person interaction (communication), most often in combination with navigation commands given by a synthetic voice, as well as control by the driver by the aid of voice commands. The primary role of the in-car navigation system is to be aware of the vehicle’s position, speed, and heading at all times to plan the journey and advise the driver to the destination.

Today’s in-car navigation systems are generally made out of three or four building blocks:

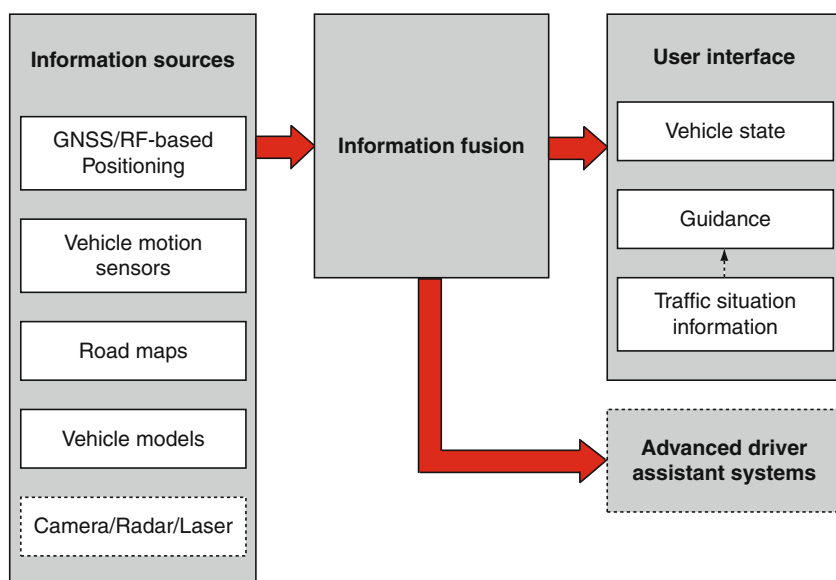
1.1 Navigation System Building Blocks

- Information source block
- Information fusion block
- User interface block
- Advanced driver-assistant system block (generally, only exists in high-end factory installed navigation systems)

These building blocks and the information flow between them are illustrated in  Fig. 17.2. In this chapter, we will focus on outlining the main ideas and the technology behind the information source block and information fusion block. We will start by looking at the information source block and the characteristics of four commonly used information sources, global navigation satellite systems (GNSSs), vehicle motion sensors, road maps, and mathematical models of the vehicle dynamics. After that we will move over to the information fusion block and look at different filter structures and algorithms



■ Fig 17.1
In-car navigation head unit (Courtesy: Pioneer Corporation)



■ Fig 17.2
Conceptional description of the building blocks of an in-car navigation system. The blocks with dashed lines are generally not a part of current in-car navigation systems but will likely be a major part of next-generation systems

for fusing and converting the information from the different information sources into a reliable navigation solution.

Before we start looking at the information source and information fusion block, the four figures of merit characterizing the performance of a navigation system are reviewed. The four performance measurements are (Skog and Händel 2009):

1.2 Navigation System Figures of Merits

- *Accuracy* – The degree of conformity of information concerning position, velocity, etc., provided by the navigation system relative to actual values.
- *Integrity* – A measure of the trust that can be put in the information from the navigation system, that is, the likelihood of undetected failures in the specified accuracy of the system.
- *Availability* – A measure of the percentage of the intended coverage area in which the navigation system works.
- *Continuity of service* – The probability of the system to continuously provide information without nonscheduled interruptions during the intended working period.


When setting up the performance specification for an in-car navigation unit, it is advisable to specify the required performance in terms of these four figures of merit. Then, when deciding on which information sources to include in the design of an in-car navigation system, the considered information source can be evaluated based on how they contribute to the overall system meeting the specified performance figures.

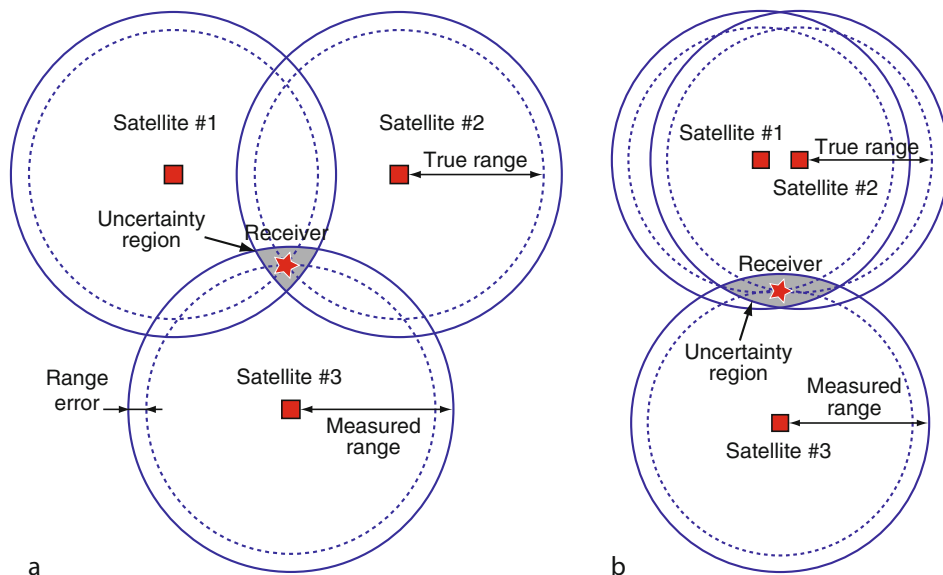
2 Global Navigation Satellite Systems

There are several GNSSs available or under construction.

2.1 Global Navigation Satellite Systems

- *GPS* – The US government maintained Global Positioning System.
- *GLONASS* – The Russian government maintained Globalnaya Navigatsionnaya Sputnikovaya Sistema.
- *Galileo* – The European Space Agency maintained Galileo system is scheduled to be fully operational by 2013.
- *COMPASS* – The Chinese government project COMPASS is a GNSS system intended to supersede the Chinese regional navigation satellite systems Beidou-1. The system is scheduled to have worldwide coverage by 2020.

The basic operational idea of a GNSS is that receivers measure the time of arrival of satellite signals and compare it to the transmission time to calculate the signals' propagation time. The time propagations are used to estimate the distances from the GNSS receiver to the satellites, so-called range estimates. From the range estimates, the GNSS receivers calculate position by means of multilateration. This is illustrated in  Fig. 17.3. The accuracy of the position estimates is dependent on both the accuracy of the range measurements and the geometry of the satellites used in the multilateration.



■ Fig 17.3

Conceptual description of the positioning of a GNSS receiver. Under ideal circumstances, the propagation times of the satellite signals calculated by the GNSS receiver correspond to the true ranges between the receiver and the satellites, and the position of the receiver is given by the interception of the circles (spheres in 3 dim). Due to errors in the range estimates, there is no single interception point, but rather an interception area (volume in 3 dim) reflecting the possible positions of the receiver. The size of the interception area (volume in 3 dim) depends both upon the size of the range errors and the geometry of the satellite constellation

The errors in the range estimates can be grouped together, depending on their spatial correlation, as common mode and non-common mode errors (Farrell and Barth 1998; Grewal et al. 2007).

2.2 GNSS Error Sources

- *Common mode errors* – Errors that are highly correlated between GNSS receivers in a local area (50–200 km) and are due to ionospheric radio signal propagation delays, satellite clock and ephemeris errors, and tropospheric radio signal propagation delays.
- *Non-common mode errors* – Errors that depend on the precise location and technical construction of the GNSS receiver and are due to multipath radio signal propagation and receiver noise.

In Table 17.1, the typical standard deviation of these errors in the ranging estimates of a single-frequency GPS receiver, working in standard precision service (SPS) mode, is

■ **Table 17.1**
Standard deviations of errors in the range measurements in a single-frequency GPS receiver (Farrell and Barth 1998)

Error Source	Standard deviation [m]
<i>Common mode</i>	
Ionospheric	7.0
Clock and ephemeris	3.6
Tropospheric	0.7
<i>Non-common mode</i>	
Multipath	0.1–3.0
Receiver noise	0.1–0.7
Total (UERE)	7.9–8.5

given (Farrell and Barth 1998). Depending on the geometry of the available satellite constellation, the error budget for the standard deviation of the user equivalent range error (UERE) can be mapped to a prediction of the corresponding horizontal position accuracy.

2.3 Horizontal Position Accuracy Prediction

The standard deviation of the UERE can together with the horizontal dilution of precision (HDOP) figure can be mapped into a prediction of the horizontal position accuracy according to

$$\text{CEP} = \sqrt{\ln 2} \times \text{HDOP} \times \text{UERE} \tag{17.1}$$

Here CEP (circular error probability) denotes the radius of a circle that contains 50% of the expected horizontal position errors. The HDOP is a figure of merit that reflects the goodness, from a multilateration accuracy point of view, of the geometry of the satellite constellation. For the GPS, the HDOP value is below 1.6 at 99.99% of the time, if all satellites above the horizon can be, and is, used in the multilateration. Since (● Eq. 17.1) is based on several underlying assumptions, the CEP figure should only be used as a rough indication of magnitude of the position error.

2.4 Differential GNSS

Since common mode errors are the same for all GNSS receivers in a restricted local area, they can be compensated by having a stationary GNSS receiver at a known location that estimates common mode errors and transmits correction information to rover GNSS receivers. This technology is commonly referred to as *differential GNSS* (DGNSS).

■ **Table 17.2**
Overview of satellite-based augmentation systems (SBASs) that regionally provide correction information for the GPS and GLONASS systems. SBASs also provide information regarding the integrity of the signals from the various satellites, and serve as additional satellites. Thereby, SBASs enhance the available satellite constellation

Coverage area	Augmentation system
North America	Wide Area Augmentation System (WAAS).
Europe	European Geostationary Navigation Overlay Service (EGNOS).
Japan	Multifunctional Satellite Augmentation System (MSAS).

The correlation of the common mode error decreases with the distance between the reference station and the rover unit. This problem can be solved by employing a network of reference stations over the intended coverage area. The errors observed by these stations are constantly sent to a central processing station, where a map of the ionospheric delay, together with ephemeris and satellite clock corrections, is calculated. The correction map is then relayed to the user terminals (GNSS receivers), which can calculate correction data for their specific location. There are several satellite-based augmentation systems (SBASs) that, through geostationary satellites, regionally provide correction information free of charge for the GPS and GLONASS systems. A list of SBABs and the geographical areas they cover can be found in ► [Table 17.2](#).

It should be pointed out that the discussion earlier in this section about performance characteristics and augmentation systems has focused on single-frequency GNSS receiver units. Using more complex receiver structures, it is possible to achieve real-time position accuracy on a decimeter level, although the required receiver units are currently far too costly for use in commercial in-car navigation systems.

2.5 GNSS Integrity Monitoring

The inherent weakness of all radio signal-based navigation methods is their reliance on information that may become erroneous, disturbed, or blocked while transferred from the external sources to the receiver. More precisely, due to malfunctions in the software or hardware of the external sources, the transmitted information may become erroneous without any notification by the receiver. Further, intentionally or not, other electronic devices may radiate radio frequency emission in the frequency spectrum used for the GNSS signals, causing interferences. Moreover, the environment surrounding the receiver may cause multipath propagation, distortion, and attenuation of the radio signal, thus complicating proper signal acquisition and distorting the information accessible by the receiver. Therefore, to create a reliable navigation system, the system should incorporate integrity monitoring. There are three key steps in integrity monitoring.

2.5.1 Key Steps in Integrity Monitoring

1. *Fault detection* – The detection of abnormalities in the information from the sensors or subsystems.
2. *Fault isolation* – The determination of which sensors or subsystems that do not work as they should and are providing false information.
3. *Fault exclusion* – The exclusion of the sensors or subsystems that are malfunctioning.

The availability of the navigation system to perform these three tasks is directly related to its integrity and continuity of service. The first task, the fault detection, provides the actual integrity monitoring by timely alert if the error in the calculated navigation solution may exceed the predefined protection levels. The second and third task, that is, the fault isolation and exclusion, enhances the continuity of service of the navigation system by isolating and excluding the faulty information before it contaminates the navigation solution.

A necessity to be able to perform integrity monitoring is the availability of redundant information in the system. For a civilian GNSS receiver, this redundant information can be obtained from an augmentation system such the EGNOS or WAAS system by observing more satellite signals than the minimum number necessary to compute a position estimate, or by a complimentary navigation system, such as dead-reckoning or inertial navigation. The inherent weakness of integrity monitoring via an augmentation system, such as an SBAS, is that its signals are also vulnerable to jamming, interference, and blockage. For GNSS receivers, self-evaluated integrity monitoring through observation of more satellite signals than necessary to compute a position estimate is referred to as receiver autonomous integrity monitoring (RAIM).

Even if the GNSS receivers' positioning accuracy and integrity is enhanced by various augmentation systems, the problems of poor satellite constellations, satellite signal blockages, and signal multipath propagation in urban environments remain. With the start-up of the Galileo and COMPASS system, the number of accessible satellites will increase and the probability of poor satellite geometry and signal blockages in urban environments will be reduced. Further, the integrity of the provided navigation solution will increase since several separate systems are available for navigation. Still, there will be areas such as tunnels where reliable GNSS receiver navigation solutions will not be available. Therefore it is necessary to use additional navigation means and aides, such as dead-reckoning, inertial navigation, or maps-matching, to produce a reliable and robust in-car navigation system.

3 Vehicle Motion Sensors

There are a number of sensors such as wheel odometers, magnetometers, accelerometers that can provide information about a vehicle's motion that may be used in combination with a GNSS receiver. These motion sensors can roughly be categorized as self-contained

sensors or non-self-contained encoder-based sensors, depending upon if they directly measure the motion of the car or if they, through an encoder, measure the motion of some moving part of the car and thereby provide information about the motion of the car. The two sensor categorized also differ in their error characteristics. Next follows a review of the most commonly used sensor in each category, and their error characteristic.

Sensors commonly used as a complement to GNSS receivers:

1. Non-self-contained encoder-based sensors
 - *Steering encoder* – Measures the front wheel direction
 - *Odometer* – Measures the traveled distance
 - *Velocity encoders* – Measures the wheel velocities (indirectly, the heading)
2. Self-contained sensors
 - *Accelerometer* – Measures the specific force (acceleration)
 - *Gyroscope* – Measures the angular rotation velocity

3.1 Non-Self-Contained Encoder-Based Motion Sensors

The three most commonly used non-self-contained encoder-based vehicle motion sensors and the quantities they measure are:

- A *steering encoder* measures the angle of the steering wheel. Hence, it provides a measure of the angle of the front wheels relative to the forward direction of the vehicle platform. Together with information on the wheel speeds of the front wheel pair, the steering angle can be used to calculate the heading rate of the vehicle.
- An *odometer* provides information on the traveled curvilinear distance of a vehicle by measuring the number of full and fractional rotations of the vehicle's wheels. This is mainly done by an encoder that outputs an integer number of pulses for each revolution of the wheel. The number of pulses during a time slot is then mapped to an estimate of the traveled distance during the time slot by multiplying by a scale factor depending on the wheel radius.
- A *velocity encoder* provides a measurement of the vehicle's velocity by observing the rotation rates of the wheels. If separate encoders are used for the left and right wheel of either the rear or front wheel pair, or if separate encoders are used for the wheels on one side of the vehicle, an estimate of the heading change of the vehicle can be found through the difference in wheel speeds. Information on the speed of the different wheels is often available through the sensors used in the antilock breaking system (ABS).

These notions of how to estimate traveled distance, velocity, and heading of the vehicle are all based on the assumption that wheel revolutions can be translated into linear displacements relative to the ground. However, there are several sources of inaccuracy in the translation of wheel encoder readings to traveled distance, velocity, and heading change of the vehicle. They are (Skog and Händel 2009):

Errors sources in non-self-contained encoder-based sensors:

1. Nonsystematic errors
 - Wheel slips
 - Uneven road surfaces
 - Skidding
2. Systematic errors
 - Changes in wheel diameter due to variations in temperature, pressure, tread wear, and speed
 - Unequal wheel diameters between the different wheels
 - Uncertainties in efficient wheelbase (track width)
 - Limited resolution and sample rate of the wheel encoders

The first three error sources are terrain dependent and occur in a nonsystematic way. This makes it difficult to predict and limit their negative effect on the accuracy of the estimated traveled distance, velocity, and heading. The four remaining error sources occur in a systematic way, and their impact on the traveled distance, velocity, and heading estimates are more easily predicted. The errors due to changes in wheel diameter, unequal wheel diameter, and uncertainties in efficient wheelbase can be reduced by including them as parameters estimated in the sensor integration.

3.2 Self-Contained Motion Sensors

The two most commonly used non-self-contained encoder-based vehicle motion sensors and the quantities they measure are:

- An *accelerometer* provides information about the acceleration of the object to which it is attached. More strictly speaking, an accelerometer produces an output proportional to the specific force exerted on the sensor projected onto the coordinate frame mechanized by the accelerometer (Britting 1971). According to the principle of equivalence, it is impossible to instantaneously distinguish between gravitational and inertial forces. Hence, the output of an accelerometer contains both forces, referred to as the specific force.
- A *gyroscope* measures the angular rotation velocity of the object relative to the inertial frame of reference.

By equipping the vehicle with inertial sensors, that is, accelerometers and gyroscopes, information about the vehicle's acceleration and rotation rate is obtained and can be mapped into estimates of the vehicle's attitude, velocity, and position. In order to measure the vehicle's dynamics in both long- and cross-track directions, a cluster of inertial sensors is needed, referred to as an inertial sensor assembly (ISA). Depending on the construction of the navigation system, the ISA may consist of solely accelerometers, but more frequently a combination of accelerometers and gyroscopes is used. In general, a six-degree-of-freedom ISA, that is, an inertial measurement unit (IMU) designed for

unconstrained navigation in three dimensions, consists of three accelerometers and three gyroscopes, where the sensitivity axes of the accelerometers are mounted to be orthogonal and span a three-dimensional space, and the gyroscopes measure the rotations around these axes.

Historically, inertial sensors have mostly been used in high-end navigation systems for missile, aircraft, and marine applications due to the high cost, size, and power consumption of the sensors. However, with the progress in micro-electromechanical-system (MEMS) sensor technology, it has become possible to construct inertial sensors meeting the cost and size demands needed for low-cost commercial electronics, such as vehicle navigation systems. However, the price paid with currently available sensors is a reduced performance characteristic. In (El-Sheimy and Niu 2007), a discussion of the usefulness of MEMS sensors in vehicle navigation and their limitations is presented. Their usefulness in navigation primarily depends on the MEMS gyroscope development.

Unlike odometers, velocity encoders, and magnetic compasses, whose errors are partly related to the terrain in which the vehicle is traveling, inertial sensors are fully self-contained. However there are several error sources associated with inertial sensors which must be considered. Some of the most significant inertial sensor errors can be categorized as (Titterton and Weston 2004; Grewal et al. 2007):

3.3 Errors Sources for Inertial Sensors

- *Biases* – The nonzero output from the sensor for a zero input
- *Scale factors* – The uncertainty in linear scaling between the input and output
- *Nonlinearities* – The uncertainty in nonlinear scaling between the input and output
- *Noise* – The random errors in the measurements

Each of these error categories, excluding the noise, in general includes some or all of the following components: fixed terms, turn-on to turn-on varying terms, random walk terms, and temperature varying terms. The fixed terms, and to a large extent the temperature varying terms, can be estimated and compensated by calibration of the sensors. Turn-on to turn-on terms differ from time to time when the sensor is turned on, but stay constant during the operation time, whereas the random walk error slowly varies over time. The sensors' turn-on to turn-on and random walk error characteristics are therefore of major concern in the choice of sensors and information fusion method. Besides the error components discussed above, there are also error components due to the inevitable imprecision in the mounting of the sensors as well as motion-dependent error components, which may be necessary to consider in the choice of sensors and information fusion algorithms.

4 Coordinate Systems

Before continuing with a discussion on how the information supplied by vehicle-mounted sensors is processed into an estimate of the vehicle's position, velocity, and attitude, or is

exchanged with the interfacing information sources in the system, it is essential to introduce a few coordinate systems. The four most frequently used coordinate systems in in-car navigation are:

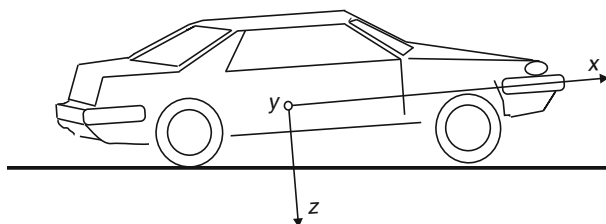
4.1 Coordinate Systems

- The vehicle coordinate system
- The earth-centered inertial (ECI) coordinate system
- The earth-centered earth-fixed (ECEF) coordinate system
- The geographic coordinate system

The *vehicle coordinate system*, sometimes referred to as the body coordinates, is the coordinate system associated with the vehicle. Commonly, but not necessarily, it has its origin at the center of gravity of the vehicle, and the coordinate axes are aligned with the forward, sideways (to the right), and downward directions associated with the vehicle, as illustrated in [Fig. 17.4](#). The information provided by vehicle-mounted sensors and the motion and dynamic constraints imposed by the vehicle model are generally expressed with reference to this coordinate system.

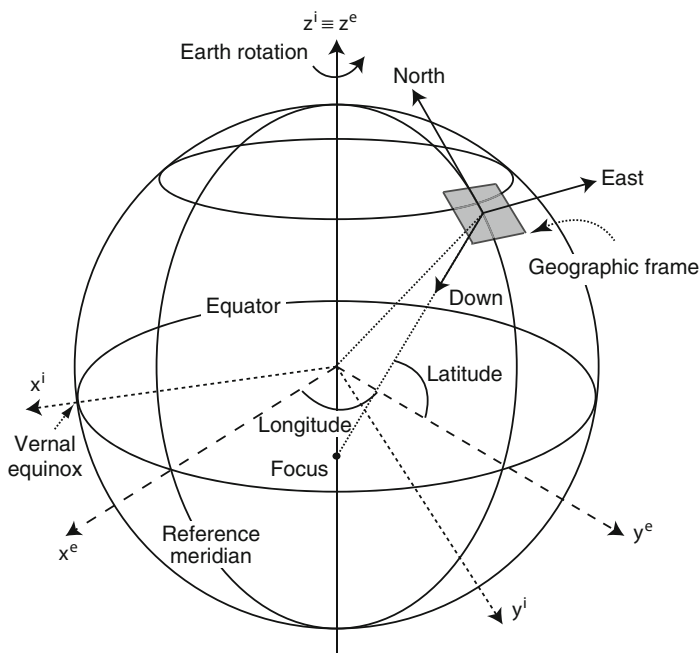
The *earth-centered inertial coordinate system* is the favored inertial coordinate system for navigation in a near-earth environment. The origin of the ECI coordinate system is located at the center of gravity of the Earth (viz., a geocentric coordinate system). Its z -axis is aligned with the spin axis of the Earth, the x -axis points toward the vernal equinox, and the y -axis completes the right-hand orthogonal coordinate system. The coordinate system is illustrated in [Fig. 17.5](#). The accelerations and angular velocities observed by the inertial sensors are measured relative to this coordinate system.

Closely related to the ECI coordinate system is the geocentric *earth-centered earth-fixed coordinate system*, which also has its origin at the center of gravity of the Earth but rotates with the Earth. Its first coordinate axis (x -axis) lies in the intersection between the primer meridian and the equator plane, the z -axis is parallel to the spin axis of the Earth,



■ Fig 17.4

The vehicle coordinate system is the coordinate system associated with the vehicle. Commonly, it has its origin at the center of gravity of the vehicle, and the coordinate axes are aligned with the forward, sideways (to the *right*), and downward directions associated with the vehicle



■ Fig 17.5

Illustration of the earth-centered earth-fixed frame (axes denoted by the superscript e), the geographic frame (axes denoted North, East, and down), and the earth-centered inertial coordinate frame (axes denoted by the superscript i)

and the y -axis completes the right-hand orthogonal coordinate system (refer to [Fig. 17.5](#) for an illustration). Nearly related to the geocentric ECEF coordinate system is the geodetic coordinate system defined by the World Geodetic System (WGS) 84 datum, commonly referred to as the geodetic ECEF coordinate system. A geodetic coordinate system representation is based on an approximation of the Earth geoid (globally or locally) by an ellipsoid that rotates around its minor axis. A location in the coordinate system is described in terms of the longitude and latitude angles measured with respect to the equatorial and meridional plane associated with the reference ellipsoid. The parameters of the reference ellipsoid, such as shape, size, orientation, etc., define the datum of the coordinate system. The coordinate system defined by WGS 84 is the coordinate system used in the GPS.

The *geographic coordinate system* is a local coordinate frame whose origin is the projection of the vehicle coordinate system origin onto the Earth's geoid. The x -axis points toward true north, the y -axis to the east, and z -axis completes the right-hand orthogonal coordinate system pointing toward the interior of the Earth perpendicular to the reference ellipsoid (Britting 1971). The coordinate system is illustrated in [Fig. 17.5](#). Note here that the z -axis does not point toward the center of the Earth but rather along the ellipsoid normally toward one of its foci. The geographic coordinate system is generally

used as a reference when expressing the velocity components of the vehicle's motion and the attitude of the vehicle platform. The vehicle attitude is commonly described by the three Euler angles – roll, pitch, and yaw – relating the vehicle and geographic coordinate systems to each other.


Detailed descriptions of the various coordinate systems used in navigation, together with common coordinate transformations, are found in the standard textbooks on inertial navigation (Britting 1971; Chatfield 1997; Farrell and Barth 1998; Rogers 2003; Titterton and Weston 2004; Grewal et al. 2007).

5 Information Processing Based on Vehicle Sensors


The processing of information from vehicle-based sensors to estimate position, velocity, and attitude is discussed in this section.

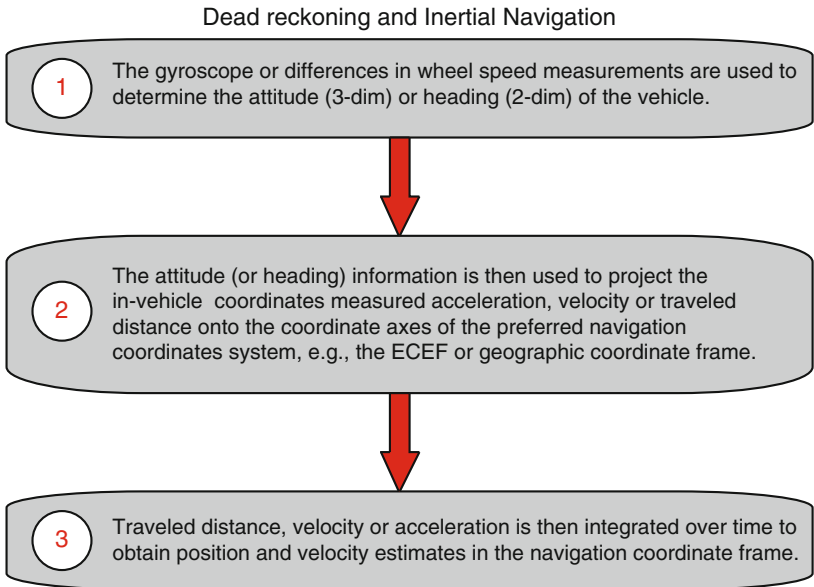
5.1 Dead-Reckoning

Velocity encoders, accelerometers, and gyroscopes all provide information on the first- or second-order derivative of the position and attitude of the vehicle. Further, the odometer gives information on the traveled distance of the navigation system. Hence, all the measurements of the discussed sensors only contain information on the relative movement of the vehicle and no absolute positioning or attitude information. The translation of these sensor measurements into position and attitude estimates will therefore be of an integrative nature requiring that the initial state of the vehicle is known, and for which measurement errors will accumulate with time or, for the odometer, with the traveled distance. Moreover, the information provided by the vehicle-mounted sensor is, except from possible fixed rotations, represented in the vehicle coordinate system. Therefore, before the sensor measurements are processed into a position, velocity, and attitude estimate, they must be transformed into a coordinate system where they are more easily interpreted, preferably the ECEF or the geographic coordinate system. Moreover, if the sensor measurements are to be used in combination with information provided by other information sources, they must be expressed in a common coordinate system.

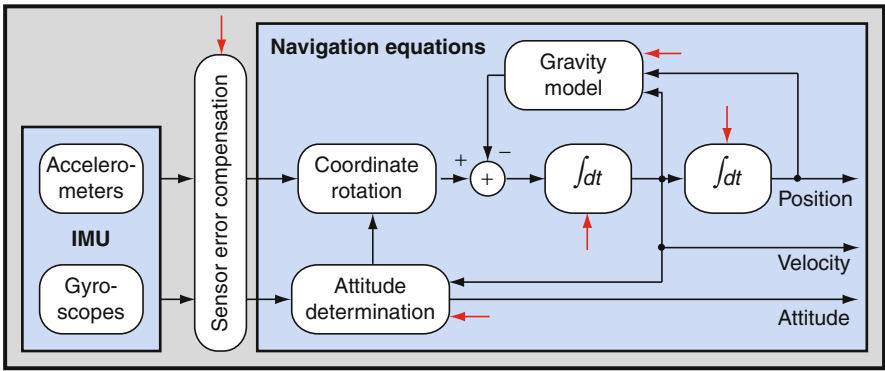
The process of transforming the measurements from the vehicle-mounted sensor into an estimate of the vehicle's position and attitude is generally referred to as dead-reckoning, or if only involving inertial sensors inertial navigation. The process of dead-reckoning (and inertial navigation) can briefly be described as in  Fig. 17.6.

5.2 Inertial Navigation

In  Fig. 17.7, a block diagram of a strap-down (The term “strap-down” refers to the fact that the gyroscopes and accelerometers are rigidly attached to the navigation platform. In a gimballed inertial navigation system, the sensors are mounted on a platform isolated



■ Fig 17.6
The process of dead-reckoning and inertial navigation



■ Fig 17.7
Conceptional sketch of a strap-down inertial navigation system. The grey arrows indicate possible points for insertion of calibration (aiding) data

from the rotations of the vehicle (Grewal et al. 2007)) inertial navigation system is shown. The inertial navigation system comprises two distinct parts: the IMU and the computational unit. The former provides information on the accelerations and angular velocities of the navigation platform relative to the inertial coordinate frame of reference, preferably the ECI coordinate system. The angular rotation rates observed by the gyroscopes are used to track the relation between vehicle coordinate system and the

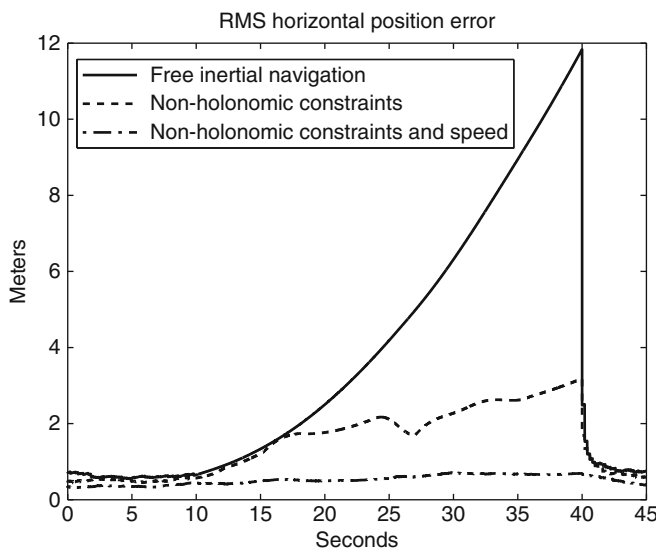
navigation coordinate frame of choice, commonly the ECEF or geographic coordinate frame. This information is then used to transform the specific force observed in vehicle coordinates into the navigation frame, where the gravity acceleration is subtracted from the observed specific force. What remains are the accelerations in the navigation coordinates. To obtain the position of the vehicle, the accelerations are integrated twice with respect to time (refer to Britting 1971; Chatfield 1997; Farrell and Barth 1998; Rogers 2003; Titterton and Weston 2004; Grewal et al. 2007 for a thorough treatment of the subject of inertial navigation).

The integrative nature of the navigation calculations in dead-reckoning and inertial navigation systems gives the systems a low-pass filter characteristic that suppresses high-frequency sensor errors but amplifies low-frequency sensor errors. This results in a position error that grows without bound as a function of the operation time or traveled distance, and where the error growth depends on the error characteristics of the sensors. In general, it holds that for a low-cost inertial navigation system, a bias in the accelerometer measurements causes error growth proportional to the square of the operation time, and a bias in the gyroscopes causes error growth proportional to the cube of the operation time (Sukkarieh et al. 1999; Dissanayake et al. 2001; Tan and Park 2005; El-Sheimy and Niu 2007). The detrimental effect of gyroscope errors on the navigation solution is due to the direct reflections of the errors on the estimated attitude. The attitude is used to calculate the current gravity force in navigation coordinates and cancel its effect on the accelerometer measurements. Since in most land vehicle applications the vehicle's accelerations are significantly smaller than the gravity acceleration, small errors in attitude may cause large errors in estimated accelerations. These errors are then accumulated in the velocity and position calculations. Hence, the error characteristics of the gyroscopes used in the IMU are of major concern when designing an inertial navigation system.

To summarize, the properties of dead-reckoning and inertial navigation systems are complimentary to those of the GNSSs and other radio-based navigation systems. These properties are:

- They are self-contained, that is, they do not rely on any external source of information that can be disturbed or blocked.
- The update rate and dynamic bandwidth of the systems are mainly set by the system's computational power and the bandwidth of the sensors.
- The integrative nature of the systems results in a position error that grows without bound as a function of the operation time or traveled distance.

Contrary to these properties, the GNSS and other radio-based navigation systems give position and velocity estimates with a bounded error but at a relatively low rate and depend on information from an external source that may be disturbed. The complimentary features of the two types of systems make their integration favorable, and if properly done results in navigation systems with higher update rates, accuracy, integrity, and ability to provide a more continuous navigation solution under various conditions and environments.



■ Fig 17.8

Empirical root-mean-square (RMS) horizontal position error growth during a 30-s satellite signal blockage in a low-cost GNSS-aided inertial navigation system. Solid curve – No constraints, Dashed curve – Non-holonomic constraints, Dashed and dotted curve – Non-holonomic constraints and speed aiding

Odometers and velocity and steering encoders have proven to be very reliable dead-reckoning sensors. For movements in a planar environment, they can provide reliable navigation solutions during several minutes of GNSS outages. However, in environments that significantly violate the assumption of a planar environment, accuracy is drastically reduced. An inertial navigation system constructed around a full-six-degree-of-freedom IMU does not include any assumption of the motion of the navigation system, and therefore is independent of the terrain in which vehicle is traveling. Moreover, it provides three-dimensional position, velocity, and attitude information. In combination with decreasing cost, power consumption, and size of the MEMS inertial sensors, this makes vehicle navigation systems incorporating MEMS IMUs attractive. However, current ultra-low-cost MEMS inertial sensors have an error characteristic causing position errors in the range of tens of meters during 30 s of stand-alone operation (El-Sheimy and Niu 2007). This is illustrated in ► Fig. 17.8 (solid curve), where the root-mean-square (RMS) horizontal position error during a 30-s GNSS signal outage in a GNSS-aided inertial navigation system is shown.

5.3 Vehicle Models and Motions

Under ideal conditions, a vehicle moving in a planar environment experiences no wheel slip and no motions in the direction perpendicular to the road surface. Thus, in vehicle

coordinates, the downward and sideways velocity components should be close to zero. This type of non-holonomic constraint can be applied to the navigation solution of a vehicle-mounted GNSS-aided inertial navigation system to reduce the position error growth during GNSS outages and to increase the attitude accuracy. In [Fig. 17.8](#) (dashed curve), the reduction in error growth using non-holonomic constraints in a GNSS-aided inertial navigation system using a MEMS IMU is illustrated. The case when observing the speed from a simulated speed encoder is also shown. In the case of both non-holonomic constraints and speed aiding (dashed and dotted curve), the error growth during the outage is negligible.

From an estimation-theoretical perspective, sensors and vehicle-model information play an equivalent role in the estimation of the vehicle state (Julier and Durrant-Whyte 2003). If there were a perfect vehicle model, such that the vehicle state could be perfectly predicted from control inputs, sensor information would be superfluous. Contrarily, if there were such things as perfect sensors, the vehicle model would provide no additional information. Neither of these extremes exists. It is clear, however, that navigation system performance can be enhanced by utilizing vehicle models. Moreover, the incorporation of a vehicle model in the navigation system may allow the use of less costly sensors without degradation in navigation performance.

There are numerous vehicle model and motion constraints, ranging from the above-mentioned non-holonomic constraints to more advanced models incorporating wheel slip, tire stiffness, etc. The literature shows that there is a lot to gain from using more refined vehicle models, especially in the accuracy of the orientation estimate. However, it is difficult to find good vehicle models independent of the driving situation. More advanced models require knowledge about several parameters such as vehicle type, tires, and environmental specifics. To adapt the model to different driving conditions, these parameters must be estimated in real time. Alternatively, the driving conditions must be detected and used to switch between different vehicle models.

6 Digital Maps and Map-Matching

Position information in terms of pure coordinates is often difficult to interpret for a driver. To assist the driver in relating the position information to a physical location, the in-car navigation system commonly displays the vehicle's position on a map. Moreover, under normal conditions, the location and trajectory of a car is restricted by the road network. Hence, a digital map of the road network can be used to impose constraints on the navigation solution of the in-car navigation system, a process referred to as map-matching.

The digital maps used in in-car navigation systems differ quite substantially from "classical" printed or digitalized maps in the way they are represented. These maps can be seen as an image built up of a set of pixels, which is easy for the human to interpret. For a computer trying to match the vehicle location on the street network and to compute

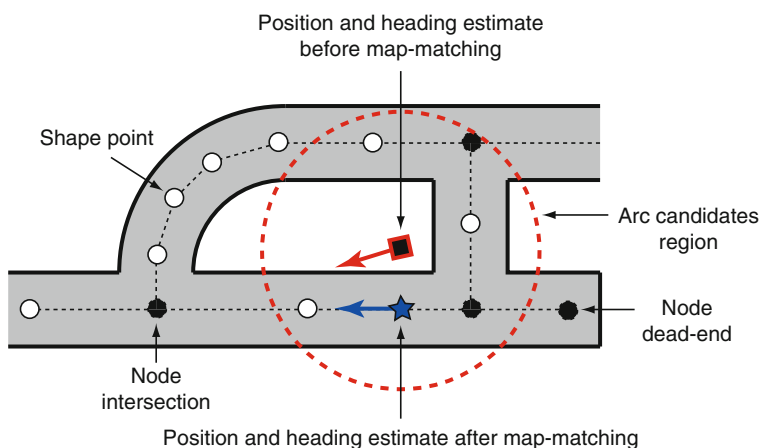
a recommended path, the image contains little information. The digital maps used in-car navigation systems and other ITS applications are therefore built up as databases containing:

6.1 Digital Maps

- *Topological and coordinates information* – connectivity properties of the features in the map and metrical information
- *Map attributes* – attributes such as road class, street name, expected driving speed, and turn restrictions

Simplified, the road network is generally represented by a planar model on digital maps, where the street system is represented by a set of arcs (i.e., curves in R^2) (Zhao 1997; Drane and Rizos 1998). Each arc represents the centerline of a road segment in the network and is commonly assumed to be piecewise linear, such that it can be described by a finite set of points (see Fig. 17.9). The first and last points in the set are referred to as nodes and the rest as shape points. The nodes describe the beginning and termination of the arc, indicating a start, dead end, or an intersection (i.e., a point where it is possible to go from one arc to another) in the street system.

Matching the output of the navigation system to the road network of the digital map generally involves three steps.



■ Fig 17.9

Road network described by a planar model. The street system is represented by a set of arcs, i.e., curves in R^2 . Generally, a set of candidate arcs/segments close to the position estimate are selected first, then the likelihood of the candidates is evaluated. Finally, the position on the most likely arc (road segment) is determined

6.2 Map-Matching Steps

1. A set of candidate arcs or segments are selected.
2. The likelihood of the candidate arcs/segments is evaluated using *geometrical* and *topological* information as well as the correlation between the trajectory history of the vehicle and the candidate paths in the map.
3. The vehicle location on the most likely road segment is determined.

The *geometrical* information includes measures such as the closeness between the position estimate and nearest road on the map, the difference in heading as indicated by the navigation system and road segments of concern, and the difference in the shape of the road segments with respect to estimated trajectory. The *topological* information criterion determines the connectivity of the candidate roads (arcs), for example, the vehicle cannot suddenly move from one road segment to another if there is no intersection point between the segments. The likelihood of the road segment candidates is found by assigning different weights to the geometrical and topological information measures and combining them. In (Quddus et al. 2007), a survey of state-of-the art map-matching algorithms is found.

Traditionally, map-matching has been a unidirectional process, where the position and trajectory estimated by the GNSS receiver, vehicle motion sensors, and vehicle-model information have been used as input to produce a position and trajectory consistent with the road network of the digital map. With improved map quality, the possibility of a bidirectional information flow in the map-matching has become feasible, viewing the map information as “observations” in the estimation of the information fusion. That is, the heading, length, etc., of the road segment identified by the map-matching algorithm can be compared to the navigation solution and used to calibrate the sensors in the system (Zhao 1997). The result is not only a higher accuracy of the position estimates but also the possibility to obtain a navigation solution when less than four satellites are available.

A series of decisions, such as map scale, map projection, and datum, are involved in the creation of a map and influence its quality and accuracy. Further, the digitalization and identification process in the construction of the spatial road network in the digital map causes additional errors. These errors can be categorized as topological or geometrical, where topological errors originate from the fact that features such as roundabouts, junctions, medians, curves, etc., of the real world have been omitted, missed, or simplified in the creation of the digital map. Geometrical errors arise from displacements of map features, such as road centerlines, junctions, etc., from their actual location. The cumulative effect of these errors influences the accuracy of the map and causes an inherent uncertainty level in the map-matching process. To establish a high level of confidence or integrity in the system, it is essential to constantly quantify reliability and look for anomalies in the result of the map-matching.

7 Information Fusion


The objective of information fusion is to obtain more information than is present in any individual information source by combining information from different sources (Yan et al. 2007). In practice, this means that by utilizing the complimentary properties of the different information sources, the information fusion tries to reduce ambiguities in the measured information and thereby expand the spatial and temporal coverage in which the system works and enhance the reliability of the system.


7.1 Filter Structures

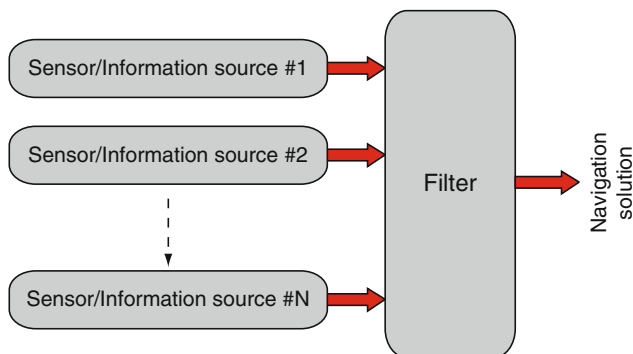
The key component in information fusion is in the concept of navigation to produce estimates of the vehicle's position, velocity, etc., from the sensor measurements. The state estimation can be realized basically in two ways: the centralized filtering mode and the decentralized filtering mode (Jekeli 2000), where the choice of mode is related to the system requirements regarding estimation accuracy, computational complexity, possibility of fault detection, and fault removal.

7.1.1 Filter Structures for Information Fusion

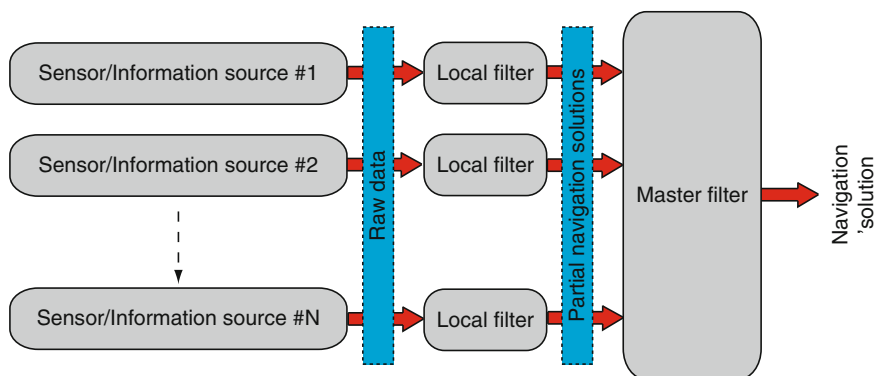
- *Centralized filtering* – The raw measurements from all information sources are processed and combined in *one* filter to produce a navigation solution.
- *Decentralized filtering* – The information from the individual subsystems are first filtered through a set of local filters working in parallel, producing a set of (partial) navigation solutions which subsequently on a periodical basis are blended by a master filter.

The centralized filtering structure is illustrated in  Fig. 17.10. The centralized filtering mode has the benefit of minimal information loss, so that all the information is directly available to the filter, which if properly designed and assigned with correct prior information theoretically should produce an optimal state estimate. However, the computational complexity of the centralized filter structure, resulting from the ability of the filter to handle all the types of raw data in the system, is often unfeasible. Also, there is the issue of robustness against spurious information from any of the information sources and how to perform fault detection and fault isolation.

The decentralized filtering structure is illustrated in  Fig. 17.11. By first filtering the information from the different subsystems through a set of local filters and then blending the information from the local filters in a master filter, the computational load in general can be significantly reduced, but at the cost of reduced estimation accuracy (Gao et al. 1993). Theoretically, as long as the correct statistics of the estimation error are propagated from the local filters to the master filter, the decentralized and centralized filter



■ Fig 17.10
Centralized filtering



■ Fig 17.11
Decentralized filtering

architecture should yield the same performance. However, error statistics are often more rigorously modeled and propagated in one filter; therefore, the centralized filter structure is often considered to offer better accuracy. One of the pros of decentralized filtering is the simplified possibility of fault detection and fault isolation and exclusion by comparing the (partial) navigation solutions of the local filters and rejecting those found erroneous.

Regardless of whether a centralized or decentralized filter structure is used, the designer of the state estimator(s) is faced with a twofold process: process and measurement dynamics modeling the choice of filtering algorithm. The modeling part is concerned with the development of a model that consistently and in a framework suitable for the state estimator (filter) describes the process and measurement dynamics of the navigation system, together with a proper description of the process and measurement noise statistics. The model must be complete enough to give an adequate description of the system and at the same time be sufficiently simple for the filtering algorithm to

become computationally feasible. The choice of filter algorithm is then a balance between computational complexity, robustness against modeling errors, and accuracy of the algorithm. Since nonlinear process and measurement models are generally used to describe information observations and vehicle dynamics, the proceeding discussion on filter algorithms will focus on nonlinear filtering methods.

7.2 Filter Algorithms

The most widely used nonlinear filtering approach, due to its simplicity, is the extended Kalman filter (EKF) in its various varieties. The idea behind the EKF is to linearize the navigation and observation equations around the current navigation solution and turn the nonlinear filtering problem into a linear problem. Assuming Gaussian distributed noise sources, the minimum mean square error (MMSE) solution to the linear problem is then provided by the Kalman filter (Kailath et al. 1999). For non-Gaussian distributed noise sources, the Kalman filter provides the linear MMSE solution to the filtering problem.

Unfortunately, the linearization in the EKF means that the original problem is transformed into an approximated problem which is solved optimally rather than approximating the solution to the correct problem. This can seriously affect the accuracy of the obtained solution or lead to divergence of the system. Therefore, in systems of a highly nonlinear nature and non-Gaussian noise sources, more refined nonlinear filtering approaches such as sigma-point filters (unscented Kalman filters), particle filters (sequential Monte Carlo methods), and exact recursive nonlinear filters, which keep the nonlinear structure of the problem, may significantly improve system performance (Daum 2005). The inherent weakness of these nonlinear filtering approaches is the curse of dimensionality. That is, the computational complexity of the filter usually grows exponentially with the dimension of the state vector being estimated (Daum 2005). Therefore, even with today's computational capacity, nonlinear filters can be unfeasible for navigation systems with high-dimensional state vectors. However, since the navigation equations in many navigation systems are only partial nonlinear, the filtering problem can be divided into a linear and nonlinear part, where the linear part, under the assumption of Gaussian-distributed noise entries, may be solved using a Kalman filter, hence reducing the computational complexity of the system (Karlsson et al. 2005; Schön et al. 2005). A short introduction to nonlinear filtering and the advantages and disadvantages of various algorithms are given in (Daum 2005).

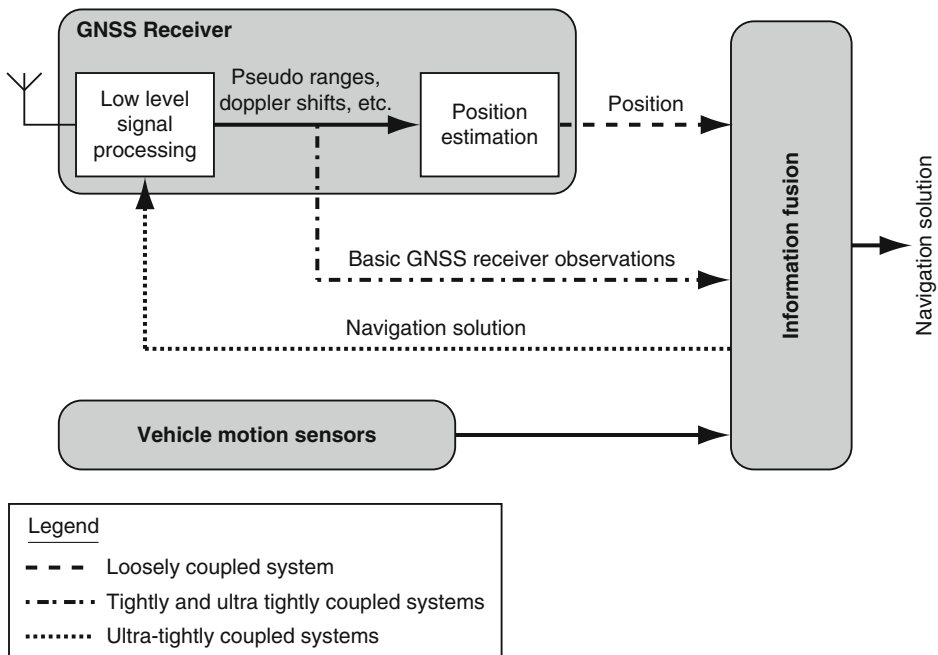
To quantify the efficiency and robustness of the chosen process model and filter algorithm, two major analytical tools are available: the posterior Cramér–Rao bound (PCRB) for the filtering problem and a sensitivity analysis. The PCRB provides a lower bound on the mean square error for the filtering problem, against which the filter algorithms are evaluated (refer to Kerr 1989; Tichavský et al. 1998 for more details). However, the mean square error lower bound provided by the PCRB is only valid when the input data to the filter is generated according to the system model used in the PCRB. To

quantify the effect of the choice of system model as well as the filtering algorithm, a sensitivity analysis may be performed (refer to Abbott and Powell 1999 for details on sensitivity analyses of Kalman filter and extended Kalman filter algorithms).

7.3 Information Fusion in Systems Including a GNSS Receiver

Without doubt, the GNSS receiver is and will remain a major component in most vehicle navigation systems, often in combination with various vehicle dynamic sensors. The information fusion between GNSS receivers and other vehicle dynamic sensors, especially inertial sensors, has therefore received a considerable amount of interest and deserves special attention.

Depending on information flows and at what level the information is exchanged and blended between the GNSS receiver and other components in the integrated system, the system architectures are commonly referred to as loosely coupled, tightly coupled, ultra-tightly coupled, or deeply integrated. (The information flow in the different couplings is illustrated in [Fig. 17.12](#).) However, no distinct bounds between the different systems exist, and the classification of integration architectures in the literature is not consistent.



■ Fig 17.12
Illustration of the information flow in a loosely, tightly, and ultra-tightly coupled system

In a *loosely coupled system* architecture, the information exchange between the GNSS receiver and other system components is unidirectional, and the navigation solution (position, velocity, etc.) of the two subsystems is blended. Generally, the information provided by the GNSS receiver is used to aid the other system components, even though the reverse is also possible. This system architecture benefits from its simplicity but has some inherent drawbacks. Due to the filter inside the GNSS receiver, the position and velocity estimates from the GNSS receiver will be correlated in space and time and with a structure most likely unknown to the information fusion algorithm designer (Farrell and Barth 1998). Thus, the (correct) statistics on the estimation error are not available for the master fusion filter, leading to suboptimal performance. Further, no information is gained from the GNSS receiver in situations when there are not enough satellite signals for the GNSS receiver to produce a navigation solution on its own.

In a *tightly coupled system* architecture, the information flow between the GNSS receiver and the other system components is still unidirectional, but instead of using the navigation solution from the GNSS receiver as input for the fusion algorithm, basic GNSS observations (pseudo-range, carrier phase, or Doppler shift estimates) are used. This has several benefits. The basic GNSS observations are not as correlated as the position and velocity solutions calculated in the GNSS receiver, and with better known statistics, leading to higher accuracy of the integrated system (Alban et al. 2003). Further, by accessing the signals from individual satellites, information is gained from the GNSS receiver even in situations when a navigation solution solely based on GNSS observations is not possible. The integrity and continuity of service of the system also benefit from the access to individual satellite signals, since fault detection at the signal level means that the individual satellite signals can be rejected instead of the full navigation solution of the GNSS receiver (Bhatti and Ochieng 2007). This is especially important for systems using low-cost vehicle sensors, which may work alone only during a shorter duration of time without the aid of the GNSS receiver.

By allowing a bidirectional information flow, so that the navigation solution of the fusion filter is fed back to aid the GNSS receiver's signal acquisition, a type of ultra-tightly coupled system architecture is obtained, also referred to as a deeply integrated system. The information feedback to the GNSS receiver is used to predict the Doppler shift in the satellite signals due to the vehicle dynamics, so that the bandwidth of the carrier tracking loops can be reduced. The benefits of this are increased accuracy of the pseudo range, carrier phase, and Doppler estimate of the GNSS receiver, and the possibility of signal acquisition at lower signal-to-noise ratios. More refined systems can at the cost of even higher computational complexity instead of the basic GNSS observations use the in-phase and quadrature phase component of the received GNSS signal as input to the fusion filter.

8 Summary

During the last decade, the in-car navigation unit has evolved from a high-price option for top-of-the-line models to a popular option at a reasonable price. On many markets, the

unit is in practice a standard utility for cars spanning from the mid-price range and upward. The main responsibility of the unit is to be aware of vehicle position at all times, to be able to plan the route and guide the driver towards the destination.

In-car navigation systems depend on information from a plurality of sensors, especially not only from global navigation satellite systems like GPS and its competitors. Fusion of information from a multitude of sensors ensures accuracy, integrity, availability, and continuity of service. The role of this chapter is to provide an overview of modern in-car navigation navigation systems, including the main information sources spanning from the satellite navigation systems, vehicle sensors, map information, and to how the fusion of the information is performed.

Many of today's in-car navigators utilize information from the traffic situation, for example, traffic message channel (TMC) information. Still, although, bidirectional communication is rare. The potential of future systems to increase traffic safety and reduce personal injuries is enormous. By cooperation between vehicles, functionalities like collision avoidance and intelligent cruise control is to be expected. By cooperative techniques, the coverage and accuracy can be improved as well. Within the European Union, some 30 MHz of spectrum has been reserved around 5.9 GHz for vehicle-to-vehicle communication. A lot of research and development is required to ensure that the future systems meet the requirements on accuracy, integrity, availability, and continuity of service, as well as other aspects like security and personal integrity. Another challenging area is the driver in the loop and how to ensure reliable man-machine and machine-man interaction, especially under stressed conditions.

One also has to remember that the main objective is typically not to guide the vehicle to its destination, but the driver. Accordingly, seamless handover between in-car and pedestrian navigation systems is an area where efforts have to be spent. Pedestrian positioning and navigation is not as mature technology as in-car positioning and navigation, by several reasons.

References

-
- Abbott E, Powell D (1999) Land-vehicle navigation using GPS. *Proc IEEE* 87(1):145–162
- Alban S, Akos DM, Rock SM (2003) Performance analysis and architectures for INS-aided GPS tracking loops. In: *Proceedings of National Technical Meeting of the Institute of Navigation, ION-NTM 03*, Santa Monica, Jan 2003
- Bhatti UI, Ochieng WY (2007) Failure modes and models for integrated GPS/INS systems. *J Navigation* 60:327–348
- Britting KR (1971) *Inertial navigation systems analysis*. Wiley Interscience, New York
- Chatfield AB (1997) *Fundamentals of high accuracy inertial navigation*. AIAA, Washington
- Daum F (2005) Nonlinear filters: Beyond the Kalman filter. *IEEE Aerospace and Electronic Systems Magazine*, vol 20, issue 8, pp 57–69
- Dissanayake G, Sukkariéh S, Nebot E, Durrant-Whyte H (2001) The aiding of a low-cost strapdown inertial measurement unit using vehicle model constraints for land vehicle applications. *IEEE Trans Robot Autom* 17(5):731–747
- Drane C, Rizos C (1998) *Positioning systems in intelligent transportation systems*. Artech House, Boston
- El-Sheimy N, Niu X (2007) The promise of MEMS to the navigation community. *InsideGPS*, March/April 2007, pp 46–56

- Farrell J, Barth M (1998) The global positioning system and inertial navigation. McGraw-Hill, New York
- Gao Y, Krakiwsky EJ, Abousalem MA (1993) Comparison and analysis of centralized, decentralized, and federated filters. *Navigation* 40(1):69–86
- Grewal MS, Weill LR, Andrews AP (2007) Global positioning systems, inertial navigation and integration, 2nd edn. Wiley, New York
- Jekeli C (2000) Inertial navigation systems with geodetic applications. Walter de Gruyter, New York
- Julier SJ, Durrant-Whyte H (2003) On the role of process models in autonomous land vehicle navigation systems. *IEEE Trans Robot Autom* 19(1):1–14
- Kailath T, Sayed AH, Hassibi B (1999) Linear estimation. Prentice Hall, Englewood Cliffs
- Karlsson R, Schön T, Gustafsson F (2005) Complexity analysis of the marginalized particle filter. *IEEE Trans Signal Process* 53(11):4408–4411
- Kerr TH (1989) Status of cr-like lower bounds for nonlinear filtering. *IEEE Trans Aeros Electron Syst* 25(5):590–601
- Quddus MA, Ochieng WY, Noland RB (2007) Current map-matching algorithms for transport applications: state-of-the art and future research directions. *Transport Res Part C EmerTechnol* 15(5):312–328
- Rogers RM (2003) Applied mathematics in integrated navigation systems. AIAA, Education Series, Reston
- Schön T, Gustafsson F, Nordlund PJ (2005) Marginalized particle filters for mixed linear/nonlinear state-space models. *IEEE Trans Signal Process* 53(7):2279–2289
- Skog I, Händel P (2009) In-car positioning and navigation technologies – a survey. *IEEE Trans Intell Transp Syst* 10:4–21
- Sukkarieh S, Nebo EM, Durrant-Whyte HF (1999) A high integrity IMU/GPS navigation loop for autonomous land vehicle applications. *IEEE Trans Robot Autom* 15(3):572–578
- Tan CW, Park S (2005) Design of accelerometer-based inertial navigation systems. *IEEE Trans Instrum Meas* 54(6):2520–2530
- Tichavský P, Muravchik CH, Nehorai A (1998) Posterior Cramér-Rao bounds for discrete-time nonlinear filtering. *IEEE Trans Signal Process* 46(5):1386–1396
- Titterton DH, Weston JL (2004) Strapdown inertial navigation technology, 2nd edn. IEE, Virginia
- Yan LP, Liu BS, Zhou DH (2007) Asynchronous multirate multisensor information fusion algorithm. *IEEE Trans Aeros Electron Syst* 43(3):1135–1146
- Zhao Y (1997) Vehicle location and navigation systems. Artech House, Boston

18 Evolution of In-Car Navigation Systems

Koichi Nagaki


Software R&D Department, Car Electronics Engineering Division,
Pioneer Corporation, Kawagoe-Shi, Saitama-Ken, Japan

1	<i>Introduction</i>	464
2	<i>Trace the History of In-Car Navigation System</i>	465
3	<i>Car Navigation System Architecture</i>	467
3.1	Conversion of In-Car Navigation System Architecture	467
3.2	Portable Navigation Device Architecture	470
4	<i>Trend and Future of In-Car Navigation System Functions</i>	473
4.1	Navigation	473
4.2	Audio and Video	475
4.3	Communication	477
4.4	Voice Recognition and Speech Synthesis	478
4.5	Display	480
4.6	Camera Application	481
4.7	Green Technology Application	482
4.8	Electric Vehicle Navigation System	484
5	<i>Conclusion</i>	486

Abstract: In-car navigation systems do not consist of only navigation functions that are combinations of GPS and Map. Present in-car navigation systems are an integrated system that mainly consists of navigation function, Audio and Video function, and communication function. This chapter provides the introduction of in-car navigation system that has various functions and connects different devices, and also shows future in-car navigation systems. Firstly, this chapter provides the basic knowledge of in-car navigation system by tracing back through the history of in-car navigation system and the system architecture. These explanations give insight into the main hardware and software components of in-car navigation systems. Further, explaining the architecture of Portable Navigation Device (PND), it shows the characteristic of PND. Secondly, this chapter briefly describes, showing trends for the future, the main software components of in-car navigation systems, i.e., navigation function, audio and video function, and communication function. Enhancement of navigation functions seems to be slowing down, but there is still considerable room for growth by linking to network. Voice recognition and speech synthesis, also covered in this chapter, would become more attractive function by linking to network. In-car navigation systems are designed to connect various devices, e.g., smart phone, portable audio device, camera, rear monitor, and ITS devices. This chapter also describes about two more devices: camera device to be connected to in-car navigation system and display device that takes center stage of front. Lastly, we look “green technology” application of navigation functions, and the in-car navigation system for electric vehicles, which functions would be different from previous in-car navigation system to provide the useful applications.

1 Introduction

Recently, In-Car Navigation System has integrated various functions, allowed the connection of different devices, and it is likely that in-car navigation system will incorporate elements of other functions. The first after market in-car navigation system with CD-ROM disk device, which had only navigation function, was launched in 1990. After that time, storage media that stores map-related information changed from CD-ROM to DVD-ROM, and later HDD. Big capacity of storage media and memory, and technical advantage of CPU made car navigation system change multiple functions and improve navigation functions to calculate a route quickly, to guide a route clearly, and to increase the location data capacity for searches. With great CPU power and graphic display ability using a graphic display controller, it has enhanced navigation map display and operation menu display. The in-car navigation display has got the best position to display the pertinent car information. Therefore, in-car navigation system has played the role of the functions of center display and operation in vehicle. Car users are familiar with audio application, this demand is high, and therefore in-car navigation systems needed to support about AV applications. Navigation unit and Audio/Video unit were separated from each other before, but now the style of all-in-one unit that integrates navigation unit with AV application has been established. Though the visual function is mainly TV and DVD-video playback, the AV function to connect digital storage media and digital music portable device to navigation

system has been enhanced recently. In-car navigation system in early days worked independently; however, recent in-car navigation systems work by communicating to the outside system through the mobile phone network. With regard to the communicating system, a “probe car” system using ITS technology has been established. The in-car navigation system in “probe” car system not only can send the data of car positioning and running speed of the car to the server of “probe car” system but also can get the road information about traffic jams, accidents, and highway regulation from this server. Amid growing international interest regarding green technology, the technology of calculating the fuel consumption and fuel-efficient vehicle routing has been introduced. In-car navigation systems for electronic vehicle have also developed to provide the services such as the indication of travel area and the location of facilities for battery charge. As mentioned above, in-car navigation systems have not only navigation functions but also other functions to connect various equipments and to provide the various services for a driver.  [Figure 18.1](#) shows the example of in-car navigation system that is connected to a number of devices.

2 Trace the History of In-Car Navigation System

Tracing the history and evolution of in-car navigation system, an inertial navigation system was developed to be used in first automotive navigation systems in 1981. For inertial navigation, the moving direction and distance of a vehicle are detected by a precise gas-rate gyro sensor and a distance sensor measuring tire-rotation (Tagami et al. 1983). The system

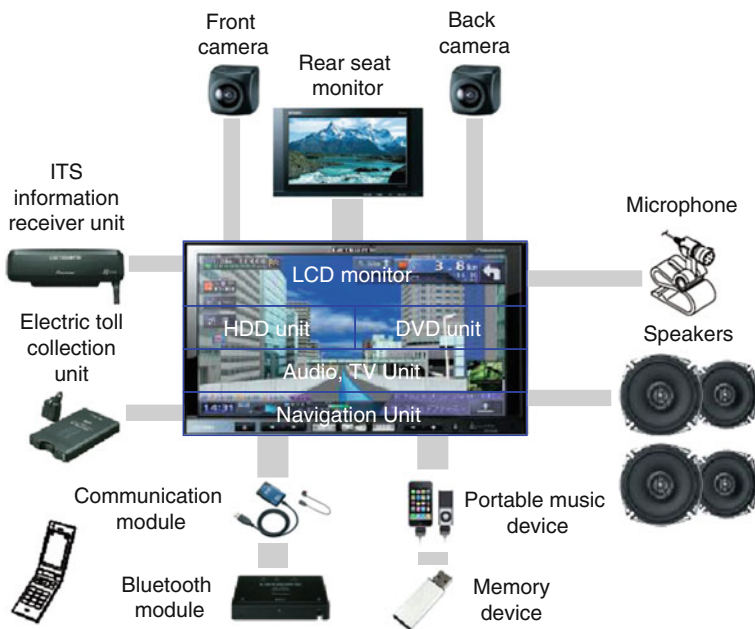


 Fig. 18.1

Example of in-car navigation system structure

displays the current position in terms of dots on a 5-in. CRT screen with a transparent map sheet. In 1987, the in-car navigation system that corresponded to the prototype of the current navigation system was developed. It used a geomagnetic sensor to perform dead reckoning (DR) and displayed digital electronic map stored in a CD-ROM system for map storage media that had large capacity and capable of random-access. In 1988, an in-car navigation system with map matching technology that was to correct current location by combining the road geometry of the digital map and the vehicle's traveled path.

In 1990, car navigation systems that used the GPS satellite signal were developed and launched into the market one after another for Car OEM and commercial market. The technology "displaying the car position on the map correctly and accurately" was established as one of the three major elements of car navigation, using DR technology based on the geomagnetic sensor and wheel speed measurements. A lot of CD-ROM systems were used for in-car navigation systems, and this became the most popular map storage medium media.

In 1991, in-car navigation to show a route to the destination was developed and mounted on the car. As a result, the function "proposing a route to the destination," another of three large elements of car navigation was achieved. In 1992, it became possible to achieve an adequate route guidance that used sound and voice interface to provide drivers turn-by-turn guidance to a destination, and display a magnified view of intersections. As the result, the third of the three basic functions of car navigation system, "guiding precisely to the destination," was achieved. Moreover, the auto re-route when coming off from the route was achieved. The basic function of the in-car navigation was established in these years, and the performance enhancement of these basic functions of the in-car navigation with a view to supporting careful and safety driving was attempted. Activity for the construction of VICS (Vehicle Information and Communication System) for acquisition of the traffic information started in 1992 in Japan (VICS 2006). The technology development of the in-car navigation system to show a traffic congestion road and directions to avoid traffic jams began. In 1996, the information service by VICS started in Japan, and the in-car navigation systems that were designed to correspond with this system were developed. A VICS receiver unit connected to in-car navigation system that received the road traffic information communicated in three ways: FM multiplex broadcasting to provide prefectural wide-area information all at once, Radio wave Beacons mainly installed on highways, and Infrared Beacons installed on major roads. The in-car navigation changed from a stand-alone system to a system connected to the outside world.

In 1997, in-car navigation systems with DVD-ROM for map storage medium were put into the market. The increase of the map data and its related information led to an increase of the number of CD-ROM disks. Using DVD-ROM disk, map data can be stored in one disk that is two-layer type and has a capacity of about 8.5 GB. The media change occurred from CD-ROM system to DVD-ROM system very rapidly. All-in-one Audio and Video in-car navigation systems that are set in two-din size were developed and quickly became a major type of in-car navigation system.

In 2001, an in-car navigation system that uses HDD for map storage medium was put in the market. The new technology of vibrations-proof and dust resistance was applied to HDDs based on Note PC technology. Furthermore, the capacity of HDD has been

increased; it was 10 GB for the first generation HDD in-car navigation system, and now a capacity of 80 GB is adopted for latest product.

In 2002, an in-car navigation system with communication module for cell-phone was introduced into the commercial market. This new technology allowed in-car navigation systems to be connected to the outside through a mobile telephone network to seek for real-time information.

In 2003, a new data upload communication system using a cellular phone that corrected the travel time information of a moving vehicle came into practical use. This system also provided the road traffic information processed by the server, which enabled the in-car navigation system to give directions to avoid traffic jams in the area where the road information by VICS was not available. This system was called “probe car system”; each vehicle played not only the role of receiver of information but also a role in gathering that information. In-car navigation system has moved to a new stage by utilizing information technology to gather various information from the running vehicle.

In 2004, the PND (Portable Navigation Device) with dedicated hardware using flash memory for map storage medium was developed and introduced into the commercial market. The new car navigation domain using flash memory for map storage was built up. This new type of product with its not expensive price created a new and large in-car navigation market domain.

In 2005, the in-car navigation system to be linked with portable music player was put on the market. At this time, in-car navigation system began the reinforcement of the cooperation with the portable music device and smart phone.

In 2009, in-car navigation systems with Blu-ray disk drive for automotive grade were put into the market. Blu-ray disk drive was used only to playback a high-definition video source, not to support the data storage function. The situation was totally different from the time when DVD system for automotive had begun to be used as storage media for in-car navigation system.

► [Figure 18.2](#) shows a simple history of in-car navigation system. The evolution of in-car navigation system had been supported by the progress of both storage medium and navigation function; these two technologies have changed in-car navigation system from just a system to navigate a driver with GPS to a comprehensive solution for the in-car environment. While the original in-car navigation system worked at a stand-alone, second stage in-car navigation used one-way communication to get the road traffic information, and the third stage in-car navigation used an interactive communication to link the Internet for the probe traffic server and search function server using a cellular phone or communication module.

3 Car Navigation System Architecture

3.1 Conversion of In-Car Navigation System Architecture

In-car navigation system architecture with DVD or HDD takes a flexible configuration, easy to connect to other equipments, e.g., a storage media unit, cellular phone, back or front

1981	• Internal navigation system with gas-rate sensor
1985	• Commercial-type navigation system that used in digital electric map
1987	• DR type car navigation system with geomagnetic sensor and CD-ROM
1988	• Developed map-matching technology
1990	• GPS type car navigation system with CD-ROM
1991	• Route planning that is showed the way to destination
1992	• Route guidance with sound and voice at turn by turn • <i>VICS system development began</i>
1996	• <i>VICS system service started</i>
1997	• DVD-type in-car navigation system • All in-one audio & video in-car navigation
1999	• Hands free telephony and data communication using cellular phone system
2000	• <i>GPS selective availability turned off</i>
2001	• HDD-type in-car navigation system
2002	• Communication module built-in type In-car navigation system
2003	• To support car probe system using cellular phone
2004	• Portable navigation device with flash memory
2005	• Linking portable music player type in-car navigation system
2006	• To support digital TV
2009	• To link Blu-Ray for automotive grade

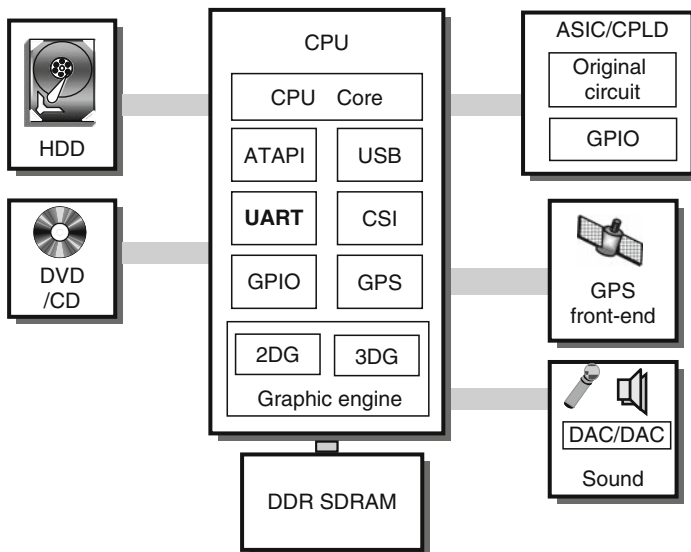
■ Fig. 18.2

In-car navigation system history

camera, information acquisition unit of traffic information, and ETC (Electronic Toll Collection) system. With regard to in-car navigation software, the in-car navigation system has to multitask drawing a map of the car position with map matching work, route planning, and turn-by-turn guidance, with real-time processing, even as it handles interruptions such as incoming calls and traffic information acquisition from broadcast or roadside devices. Therefore, most in-car navigation systems adopt a 32 bit RISC processor, graphic processor with 2D and 3D engine to draw the various types of sophisticated navigation map, and a real time OS. This in-car navigation system consists of three major blocks, i.e., CPU, GDC (Graphic Display controller), and ASIC. The ASIC needs to have many peripheral functions and to have interfaces to other peripheral devices such as HDD, DVD/CD, cellular phone, GPS, DSP, microphone, and so on. Requiring the performance of smooth map scrolling and refreshing the map display of vehicle location several times a second, the program load of drawing a navigation map is very high because navigation maps involve a large volume of small drawing objects such as roads and rectangular geometry. Therefore, a GDC with a 2D engine is used in-car navigation system to reduce CPU power needed to draw a map. In the navigation system, the GDC chip is required to draw a display without handling by the CPU. The method by which this is achieved is that the command sequence of screen structure is written as the command list on the CPU side, and this command list is sent to the GDC to be executed directly by GDC. As the drawing contents of a contiguous region in terms of time are build mostly of common graphic objects, the system stores drawing content as a display list.

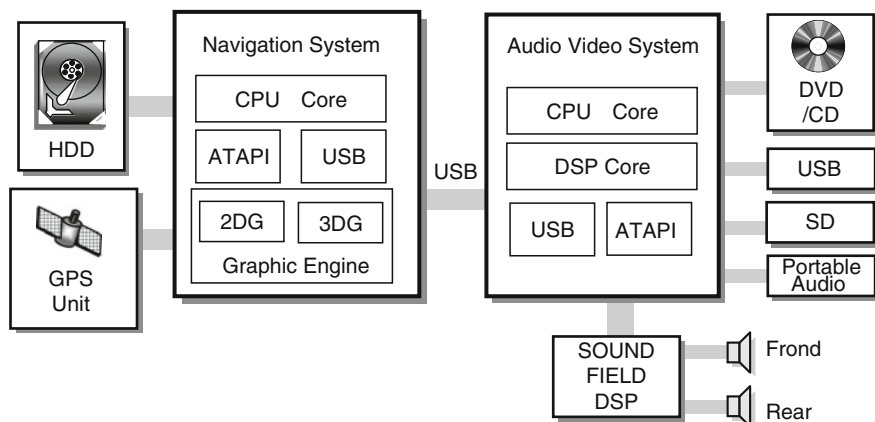
Reusing this list allows reduction of CPU load and improvement of the drawing performance as a whole. Adding devices to display a “bird’s-eye view” map and 3D landmark symbol of famous location, a GDC with 2D and 3D engines, with geometry transformation engine and shading function using a “Z buffer” method has been used. Moreover, the function of “superposed display” is available for the in-car navigation graphic chip. This function enables to display guidance and messages by “multilayer drawing technology” that supports 4–8 layers of drawing screen and “alpha blending technology” that changes transmittance in every layer. Anti-aliasing is also required to reduce rough edges around the lines of roads because of heavy volume of road line drawing in navigation map displays. In-car navigation system with HDD is needed to be equipped with an audio function to encode and decode compressed audio, to exploit the fact that HDD is rewritable. An external dedicated DSP suitable for compress audio is used because the real-time processing of compressed audio is required simultaneously to other in-car navigation functions.

As silicon process of LSI manufacture evolves, an in-car navigation core engine, with CPU, GDC, and peripheral function contained in one ASIC has been developed, and this type of core engine with CPU has been adopted in-car navigation system. ● Figure 18.3 shows the architecture of this type in-car navigation system. The feature of this type architecture is that GDC is included in-car navigation core engine LSI. As the previous GDC of the in-car navigation system existed independently, external RAM memory was individually required to work with screen plane and the working area for drawing. Graphics memory made an impact on the circuit board size and product cost. The integrated core chip where



■ Fig. 18.3

System structure of integrated in-car navigation system



■ Fig. 18.4

System structure of separated audio and video in-car navigation system

the CPU and the GDC were unified allowed for one memory as shared memory of CPU work memory and graphic memory to be adopted into the system.

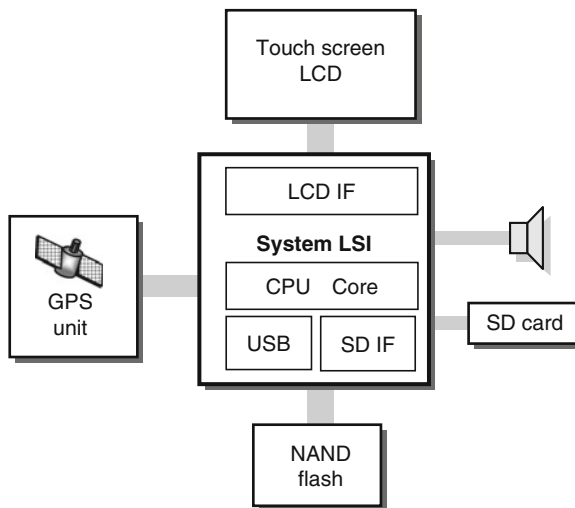
In recent years, Audio and Video function needs to play various type of audio format and video format from external storage devices, and demands to display Digital TV broadcasting has become very strong. Figure 18.4 shows the system architecture that has a separated navigation part and audio and video part to meet the requirements above. This architecture type would be adopted into in-car navigation system that requires high performance, but an in-car navigation system whose cost requirement is very strict would be composed of only a navigation part platform; in this case, the audio function would be available, but video function would be not. This is because high-frequency CPU made it possible to decode compressed audio, but decoding highly compressed video requires a dedicated DSP. There is no doubt that an in-car navigation system using core LSI with multi (2 or 4) core CPU, 3D graphic engine with more powerful performance, and video interface of output and input would be launched into the market someday. Moreover, the integrated telecommunication chip with Bluetooth, Wi-Fi, and GPS function will be connected to the in-car navigation core LSI. This architecture would be similar the smart phone system. It would eliminate barriers between the car navigation platform and the smart phone platform. This means that the in-car navigation system architecture would advance to become the standard platform architecture like a smart phone platform, and navigation core chip would be replaced by a smart phone application processor.

3.2 Portable Navigation Device Architecture

In 2002, a navigation system was developed, whose software was installed as a dedicated navigation application on a PDA (Personal Digital Assistant) device and which was the prototype of present PND (Portable Navigation Device). In 2003, a PND with dedicated

hardware was developed and launched into the market. In-dash and on-dash types of in-car navigation were mainstream systems at the time, and it could not be easily detached from a car. This type of in-car navigation system employed DVD or HDD, which is a mechanical device for map storage medium, though PND use large Flash Memory. One characteristic of developmental regime of PND is to be able to develop the navigation software without the restriction of hardware, which is totally different from the previous in-car navigation developmental style. One factor which gave better performance to PND devices to bring them up to practical level was the turning off the SA (Selective Availability) signal provided by the state department of USA. Moreover, the evolution of SoC (System on Chip) technology, stepping into deep submicron range, made it possible to integrate new technologies into one chip, with high-specification CPU core and the peripheral functions necessary for PND system. Especially, the improvement of GPS function, such as reception sensitivity and acquisition time, played an important role in bringing PND up to a practical level. ➤ [Figure 18.5](#) shows the basic system configuration diagram of PND.

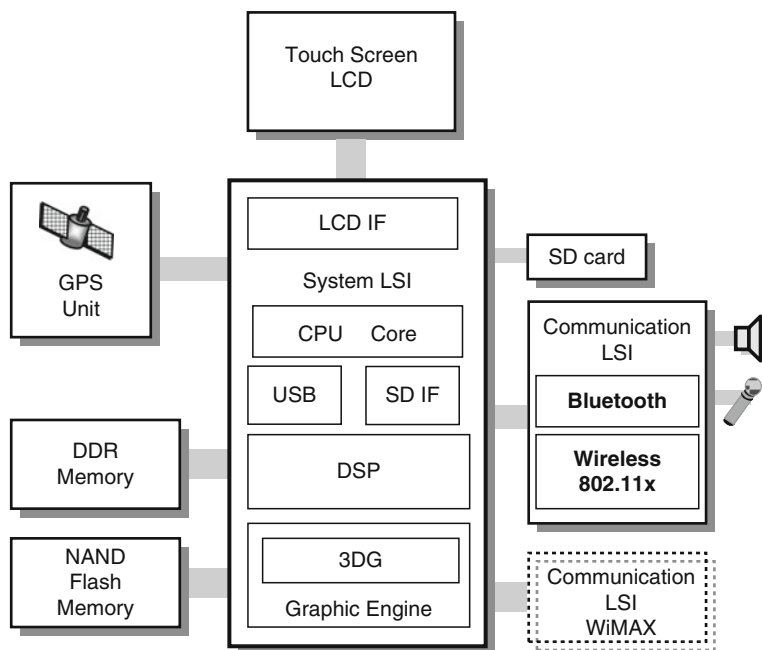
Basically, the hardware architecture is composed by the integrated SoC with CPU core, GPS chip, flash memory connected to SoC, and a Display device. It has a simpler structure and fewer components, compared to an in-car navigation system. As the number of pixel for PND is not so big, 3–4 in. screens being mainstream, a GDC unit is not required; a CPU that works at very high frequency can draw navigation map for PND. By using flash memory, the power consumption is lower than previous system, and so a PND system with a battery becomes available. As it can use the suspend-resume function that is supported by the OS, the navigation map can be displayed quickly and GPS module can find its actual location by holding the information of GPS satellite orbit.



■ Fig. 18.5
System structure of portable navigation device

Formally, the method of vehicle positioning in most PND systems was to use only GPS; but some recent PND systems work with a gyroscope to locate more correctly just like conventional in-car navigation system. The difference between two devices is that PND mainly use GPS to locate the current position, and sensors are auxiliary measures; on the other hand, a conventional in-car navigation system mainly uses sensors, and GPS is an auxiliary measure. Due to tightening of the legal regulation of operating cellular phones while driving, and expanding market of cellular phone with HFP (Hands-Free Profile)/HSP (Headset Profile) using Bluetooth technology, there are a lot of PND device with Bluetooth function to link to a cellular phone. Some chips used in PND devices have the function of the echo cancellation and noise-cancellation technology. A PND device has not only the fundamental function such as navigation but also multimedia application, performed with software-based signal processing by CPU, such as Audio/Video decoder and photo viewer, and with hardware signal processing by a dedicated chip such as the Digital TV function that is sometimes used. PNDs have reinforced the functions of the communication over the Internet using DUN profile of Bluetooth or another communication module. Also, the communication tools like on-line-search system and the road traffic information system have been reinforced to link the system to the outside world, allowing a form of “cloud computing system” on the Internet.

➤ *Figure 18.6* shows the advanced PND system configuration diagram. It would be similar to a smart phone platform that incorporates with the elements of the communication



■ Fig. 18.6

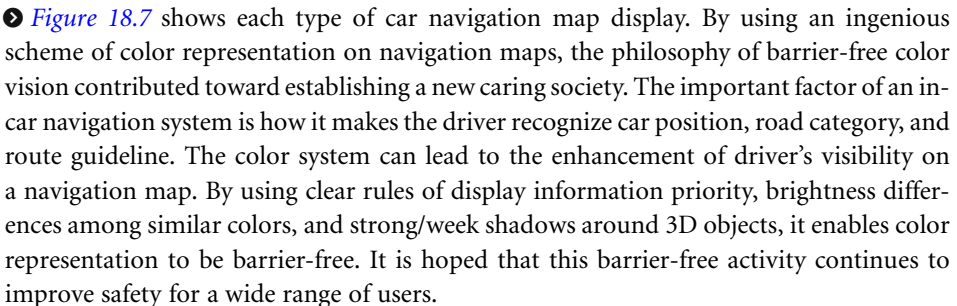
System structure of advanced PND

application. The difference is only the display size; therefore, these platforms could integrate each other, and the same SoC chip could be used in these systems that would be improved to more powerful device. If PDA gets new function to link the in-car system using vehicle network, it may get the same capabilities as, or more than, in-car navigation system.

4 Trend and Future of In-Car Navigation System Functions

4.1 Navigation

As mentioned in the previous section on car navigation history, the basic functions of car navigation consists of three major parts, which are: matching car location on a map, route planning for destinations, and route guidance. The accuracy of map matching technology has been advancing by means of sensor technology that is supported by gyro sensor and acceleration sensor (G sensor). Most of in-car navigation systems get speed pulse and reverse signal from vehicle, and the car location is pinpointed by dead reckoning of combination using GPS and these sensors which are automatically calibrated. Therefore, this dead reckoning allows in-car navigation system to have high accuracy over long-distance surveying when the car is running in tunnels, parking lots, underneath elevated railway tracks, or narrow space between tall buildings. Competition to improve “how much shorter time it takes to finish calculating a route after a destination point is set” has made in-car navigation system much more effective by improving the map data with route profile data. Nowadays, the quality of route planning and the proposal of several route plans with route options such as distance priority, high way (main load) priority, use of toll roads as a priority, not to use toll road priority, and so on are more important. Recently, recommendation of the bypass route for avoiding traffic jams, traffic control, or accidents makes getting the road traffic information more valuable. The route guidance includes the function such as “turn-by-turn with voice,” “enlarged illustration display,” and “lane display.” Especially, the function “voice guidance of turn by turn” considers both car speed and the required time to finish the voice guidance by a prescribed position before the next turn.

The evolution of GDC and map data has led to a different kind of car navigation map.  **Figure 18.7** shows each type of car navigation map display. By using an ingenious scheme of color representation on navigation maps, the philosophy of barrier-free color vision contributed toward establishing a new caring society. The important factor of an in-car navigation system is how it makes the driver recognize car position, road category, and route guideline. The color system can lead to the enhancement of driver's visibility on a navigation map. By using clear rules of display information priority, brightness differences among similar colors, and strong/weak shadows around 3D objects, it enables color representation to be barrier-free. It is hoped that this barrier-free activity continues to improve safety for a wide range of users.

The search function, that provides the destination and surrounding of car's location, is another important factor of in-car navigation system. Search technology has depended on the



■ Fig. 18.7

Example of in-car navigation map image

map data stored in a storage device like DVD or HDD. Having strengthened search function and the management of database of various types of facilities, it has been available to search by facility name, telephone number of facility and home, every category used in “yellow pages” directories, and so on. Previously, the data from local database and local search engine had been used for the search function of an in-car navigation system. However, there are problems of freshness of facility information in such data. Recently, using embedded data base technology, two new features become available: one is to use refined search technology made by name or type, and the other is to update the search database from an outside storage device, from which data is acquired by accessing the update server of the car navigation manufacture’s site, without waiting until the physical map data update service. This embedded database technology will expand to be used in-car navigation systems that provide a user-friendly searching interface to find out the location of facility. As one feature of next generation, it is expected to combine the network search function using the server search engine and the enormous quantity of data in various servers with the conventional local search function. Moreover, combining the technology of a speech recognition engine at the server side, without touching any screen keys, sending a voice data to the server, it would make it possible to input destination or surrounding facilities that the user would like to see by connecting communication device through the network. The speech recognition technology, that makes users of in-car navigation system enable to operate it without touching any screen key, can also become one of the “killer” applications. A stand-alone version type of this function has already implemented in some in-car navigation products. Moreover, the technology using the server on the Internet is going to be a successor of this function as it would be

able to provide the latest information to the users. As shown above, it is clear that there is still room for improvement of search function linking to the search sites on the Internet.

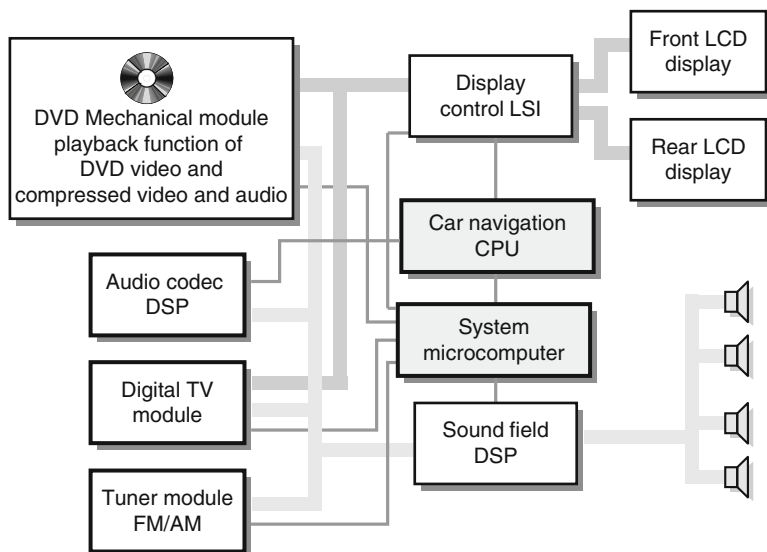
The KIWI format, including display map data, graphic data, index data, road category data, private data, and so on, has been adopted as standard map data format for in-car navigation systems with CD-ROM and DVD-ROM (Fujimoto 2001). However, a new type of map data format has been developed recently by demanding more information and less volume. This new compressed data format about 2–4 GB size will comprise the major portion of map format of an in-car navigation system, in either flash memory or SSD (Solid State Drive) up to 32 GB size. Also, new data-updating technology for map data of car navigation is to be introduced. This new technology realizes a “partial data-updating system,” and gives users the benefits of getting the latest map data, including new-opened roads without taking a long time to get data from the update server that is set up by navigation maker or automotive company. Even a map data is managed to keep the map data fresh by linking the Internet.

As described above, car navigation function in these days works not as stand-alone, but with network functions and services; linking to the outside world would result in a drastic improvement that provides more convenience for users.

4.2 Audio and Video

For in-car equipment, the percentage of vehicles with a playback function for audio and video content is high. It began with radio broadcast service reception and playback, television broadcast service, and it expanded playback of recorded media such as CD-disk, DVD-disk, or USB memory like flash memory. Even in-car navigation systems are no exception, there has been an increased tendency toward all-in-one units with Audio and Video function since the advent of DVD based systems. Now, in-car navigation systems are an all-in-one unit of Audio, Video, and Navigation. ➤ *Figure 18.8* shows the AV function structure of an in-car navigation system. The system microcomputer controls FM and AM receiver, DVD mechanical module, operation of AV source switching, Sound Field DSP, and peripheral devices related to the AV function, for example, AD/DA, electric volume, and so on. The video source that is outputted from various units is selected and sent to the latter block which is a picture processing LSI that treats screen control, adjust brightness, image compensation, and so on. The selected audio source is sent to the latter block which is a sound field DSP that treats frequency compensation in the car and making a sound field over several speaker channels. As better sound quality is an important factor in an integrated in-car navigation system with audio and video, in-car navigation systems that support technology for improving the sound quality, for example, ANC (Anti Noise Control), filtering out engine noise or road noise, would be developed in the future.

In in-car navigation system, DVD mechanical module has control of not only the playback of DVD-Video and other video source disk like CD-Video, but also playback of the various types of audio contents, MP3, AAC, WMA, and so on, that are stored on disk.



■ Fig. 18.8

Audio and video structure of in-car navigation system

Therefore, DVD Mechanical module has an AV LSI that enables it to playback several music formats. In recent systems, there is mechanical module that only supports ATAPI interface not having an AV LSI. In this case, mechanical module has responsibility for reading data. The navigation unit has a dedicated AV LSI, and playback DVD-Video, or other video-encoded video and music source, is decoded here. Because it is easier functionally to decode the AV source here than in mechanical module, and it is often the case that an in-car navigation system can decode various type of encoded video and audio source than it could before. Moreover, in low-cost models of in-car navigation systems, the in-car navigation system sometimes takes approach of having the navigation CPU perform encoding and decoding music. Especially, in the case of inexpensive in-car navigation system, not using a dedicate AV LSI, the CPU that processes the navigation functions also takes charge of AV decoding from flash memory that contains compressed music and video. From when HDD devices were adopted in in-car navigating systems, to make use of the ability to write the data to the HDD, applications where encoded music data that is converted by dedicated DSP is automatically stored in the HDD when a CD is inserted in in-car navigation system became mainstream. In-car HDD navigation systems take into account copyright protection. The music data in HDD cannot be moved to another device and the music data exists only in the HDD. Recently, in-car navigation systems which use flash memory for map and related data support the ripping function from CD to flash memory that is contained in navigation unit. The tendency to increase the amount of flash memory will continue, and the memory area that is not used for map data will also increase. Therefore, this situation would cause an increase in case of ripping from CD in in-car navigation with flash memory for map data.

Recoding media used in vehicles has evolved not only for in-car application, but use case of in-car media has been affected by the circumstances of distribution channels of content, and at-home viewing and listening. The tendency to listen to music that people “rip” using a PC or get from a network is strong recently. In-car navigation systems need to support a linking function for portable audio devices and cellular phones with BT audio to operate from in-car navigating system touch display. From now on, the applications and functions for connecting to other devices which have audio content by using a network, rather than the audio and video application that the in-car navigation performs itself, will be more important and become mainstream. Especially, it is easily expected that smart phone will evolve enormously to become a big devices, and its applications would expand to link the network world that is known “Cloud World” and business opportunities for portable audio and video will be growing. People who have smart phone would hope that they can connect of their smart phone to their in-car navigation system to operate it through the in-car navigation system display to process the content of their smart phone, and not to operate it via the smart phone display. This link application would be more important than any other application in-car navigation system has.

4.3 Communication

In-car navigation system is said to be a kind of information device to add the location information on top of map information and to provide route information adding driver's destination. The basic value of an information service is said to be information freshness, and information about traffic jams is one type of very fresh information for in-car navigation systems. Car navigation systems have a strong association with traffic information services. Accordingly, the VICS system was developed to provide traffic information to in-car navigation system in Japan. VICS services are traffic restrictions or parking availability, transmitted using three methods, i.e., FM multiplex broadcasting, radio wave beacon, and infrared beacon. By getting the VICS information, the in-car navigation system progressed to the next stage, to get the traffic information from the outside world, as opposed to previously being in-car navigation system with local information data which worked as a standalone system. Now, although the service configuration is different each country, traffic information services are provided to in-car navigations now. The communication function has been growing since in-car navigation systems that could link with cellular phones were developed in the late 1990s. Linking cellular phone by wire to in-car navigation systems, the hands-free telephony capability, and data communication and Internet access have become available. Hands-free telephony capability was achieved by using a dedicated echo cancellation device, but experiencing price competition, this function was integrated into the software that is executed by the CPU. Recently, it is often the case that hands-free telephony function is achieved by an LSI with Bluetooth control of HFP (Hands-Free Profile) and HSP (Headset Profile) because cellular phones with Bluetooth have become common. Data communication is also supported by such Bluetooth LSIs with the DUN (Dial-Up Networking) profile. As a hands-free telephony function by Bluetooth is a strong user needs, there is little doubt but that linking in-car

navigation system to connect to cellular phones using Bluetooth connections is a major connection mode. The Wi-Fi function has been attracting a lot of attention as a means of wireless communication in recent years. Wireless communication routers have become popular, and people who have portable devices or notebook PCs have had a greater opportunity to link to the Internet using such wireless communication routers. Cellular phones or communication modules can be linked to in-car navigation systems, but as a means of net wireless communication, in-car navigation systems with Wi-Fi devices are expected to develop as one means of linking devices link to the Internet. The current major application in in-car navigation systems using a communication function is Prove Car System. The Prove Car System is to gather information from running vehicles that is used to resemble a “probe.” This vehicle is called “Prove Car.” An in-car navigation system with this function records the driving history that includes vehicle location, vehicle speed, sensor information, destination, search location point, and so on. This driving history data is uploaded to the server that calculates road traffic, and present or predicted road traffic situation is sent to the vehicle so that it can avoid traffic jams. Meanwhile, the various sensors that are set up on a road gather data on the road traffic situation and traffic jam information is delivered to in-car navigation system, but facilities for road sensors are not set up for all roads. Using prove data the in-car navigation system collects, traffic jam information of roads that were not supported until now is made available over the Prove Car System. As the in-car navigation system receives traffic information from the server from the prove car system over the network link with a cellular phone or communication module, it can estimate accurately the arrival time and calculate the route to avoid traffic jams. Moreover, the data collected from prove cars is expected to be used as basic data for traffic jam prediction, and the technology of accurate traffic jam prediction would improve more. Furthermore, it is difficult to grasp the present situation from the basic infrastructure quickly, for example, weather or road surface condition, but if this information can be got from a collection of probe cars, this prove system can facilitate the collection of this environment data. The prove system is one of ITS solutions, and various utilization methods are expected to estimate traffic flow for road administration operations. Thus, the in-car navigation system is expected to play a significant role in ITS solution and growth. There is much content and services in Internet world. The opportunity to connect to outside world using Internet in in-car navigation system has been increasing. It goes without saying that a mobile information terminal is a two-way communication system. There has been a big change in having a greater opportunity to up-link to Wide Area Network from a vehicle. In-car navigation system would be expected to become gateway machine to connect the outside world to the vehicle.

4.4 Voice Recognition and Speech Synthesis

The voice recognition function has come to be used in in-car navigation system with other systems because of the particular car environment. An in-car navigation system has a high-performance CPU, enough memory to work with heavy applications, and

microphone for hands-free telephony. There is strong requirement to be able to eliminate the need for drivers to take their hands off the wheel or to take their eyes off the road for their safety. The driver can use all functions while the vehicle is stopped, but the in-car navigation system sets limits on operation while the vehicle is moving by legal requirement or self-regulation. Using the voice recognition function, the user can operate the functions safely without touching the display menu even if the vehicle is moving. There are various ways to detect whether the vehicle is moving, using the parking break, speed pulse, or GPS. Viewed in another way, the user must press the menu list in a sequence of several steps to get the intended operation, so in-car navigation operation menus have hierarchic structure that exist in, for example, the Set Destination menu of name, address, genre, telephone number, and so on. Using the voice recognition function, a single voice command can achieve the operation. For these reasons, voice recognition is seen to be of great benefit on in-car navigation system. The voice recognition engine for the in-car navigation system has a function to filter out the car noise, for example, engine noise, road noise, wind noise, and so on. A practical recognition rate about 80–90% or more is required. The main use of the voice recognition application for in-car navigation system is setting up with a “command word,” for example, “scale change,” “map view change,” “source change,” and so on. Command word recognition uses “word spotting technology,” that is, to recognize only the command word if any additional phrase exists.

A large vocabulary recognition dictionary is required to determine the destination in, for example, an address name or a facility name. Recently, voice recognition applications have expanded to include music title search, singer search, and so on, to provide more usability for portable music player connected to the in-car navigation system. The vocabularies of voice recognition dictionary used in in-car navigation system are limited by size restraints; however, it is possible to load a large size of vocabulary dictionary into a HDD car navigation system. There is also the problem of data fleshiness, for example, a data of facilities information. It is difficult to refresh vocabulary dictionary in a stand-alone type in-car navigation system, so the in-car navigation system search functions using the voice recognition system located on the server side, which holds various type of information about shops, restaurants, facility information, and so on, is expected to become more popular.

The function of speech synthesis is mainly used in route guidance at an intersection, to show a driver to a destination in safely. It is difficult for HDD in-car navigation system to continuously refresh the vast amount voice data for place-names that has been recorded by the professional narrator and must support multiple languages. It is very expensive to recode and maintain the data; therefore, speech synthesis technology is useful for in-car navigation systems. It is also difficult to develop the function so that it understands the meaning of content pronounced by user on the in-car navigation system side. However, the technology to implement voice recognition and voice synthesis on the server side has advanced recently. The natural language dialogue system in in-car navigation system, using a cellular phone to link to the server and transfer the various information which the user of car navigation requires, is expected to be available in near future. This dialogue technology, linking to the outside with network, enhances this function and would make in-car navigation system more attractive to the user.

4.5 Display

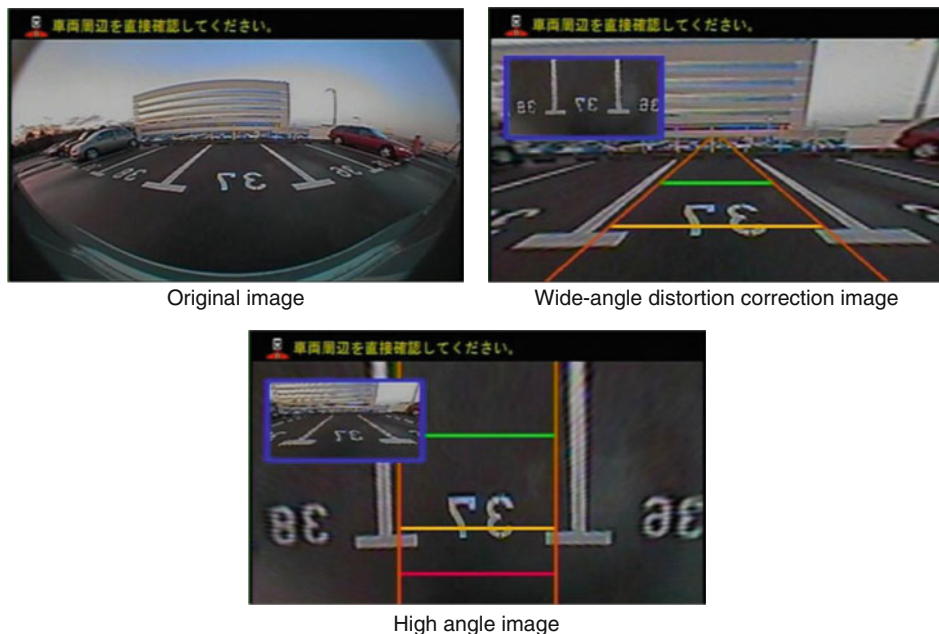
Most displays in in-car navigation system is liquid crystal; previous in-car navigation system used Wide Quarter VGA (320×240), but recent in-car navigation displays are of the Wide VGA (800×480) type with a size of about 6–7 in. Wide QVGA, which is about 3–4 in., is mostly used in PND; however, some PNDs have a display size of WVGA. In-car navigation systems with WXGA ($1,280 \times 720$) size displays exist in the market, but their number and model types are few. A touch-sensitive panel for easy operation, using a resistive touch screen, is used in a great number of car navigation systems. Some in-car navigation systems, where the display is placed where it is hard to reach, especially in OEMs in-car navigation systems, are operated by a JOG dial with a rotary encoder. The specification of liquid crystal displays for in-car navigation system is different to that for personal computers. A guaranteed operating temperature range of -30° to 85° , and super luminosity of 500 cd/m^2 in order to be able to recognize the map and letters even when subjected to direct sunlight are required to adapt the system requirements to in-car environment. An in-car navigation display also requires a 30° wide viewing angle because of being installed in the center of the front panel in the car. The response time of liquid crystal goes down under low temperature, especially falling greatly below 0°C , and the image of LCD becomes dim under such conditions. LCD for in-car navigation systems are specifically designed to improve the response time of liquid crystal to be able to see the lettering of navigation map while map is scrolling and an image from a rear view camera with bright clarity. Some in-car navigation systems have a DVD/TV playback function and can display both DVD/TV contents and navigation picture simultaneously. However while the vehicle is moving, the front display should not show DVD/TV contents in accordance with regional laws and regulations. An in-car navigation system has individual video signal circuits for video contents, to output a video signal for the dedicated rear-monitor which has only video input, but does not have TV tuner. The size of a rear monitors varies from wide 7 to 11 in. Their installation position is either the headrest of the driver's front passenger seat (embedded type) or suspended from the inner-roof (ceiling type). In-car navigation systems provide the multifunction of an audio and video system that user can enjoy in both the front and rear seats. An in-car navigation system that is connected to a Blu-ray video disk unit and can send a high-definition image signal to a rear monitor is required to have copyright protection technology. Therefore, the 1394 automotive format or HDMI interface as those widely used in home and PC implementations would be used to connect to the rear monitor. A "Head-up" display has been introduced as an optional extra. It has high affinity with the in-car navigation system because driver can confirm the navigation information easily without taking their eyes off the road as it is displayed on front window. This system displays monochrome images and is still expensive yet, but a full-color "head up" display using laser-module has been developed. As for linking between a head-up display unit and in-car navigation system, it would enable the display of information on the front window of a vehicle, such as guidance at intersection or traffic lane, announcement of the spot where a traffic accident has occurred, and a warning when the system recognizes danger, synchronizing with an

in-car camera. The display that is set up in the front center of instrument panel plays a big role as an infotainment system. Equipment in the vehicle has been computerized and digitalized rapidly, and information from inside and outside of the car has been increasing. There is a limit to the expression of necessary information for drivers with mechanical instrument panel. Thus, this vehicle information and related information would be displayed on the in-car navigation's display as the in-car navigation system has great expressive power for drawing pictures using the 2D or 3D graphic engine in the integrated SoC chip, and it can be connected to various equipments in the vehicle with comparative ease. The in-car display, when it is installed the front center of instrument panel, may serve as the in-car navigation's display or may be a display unit with great power in its CPU and GDC engine, and might evolve to be a multifunction display device that can display information of various equipments for better visibility and easier operation of these equipments.

4.6 Camera Application

Camera applications of in-car navigation system have been mainly back camera applications, where the navigation screen displays a rear side view when the gear is shifted into reverse. The in-car navigation system displays mirror image that is converted by the hardware in camera module for better visibility. Requirements for back camera include wide angle, dynamic range of brightness (e.g., in the darkness, under the sunlight, and so on), and good performance even in harsh automotive environment. This back camera application has advanced because of increasing requirements for both safety and security. Recent back camera applications have various functions enabled by digital image processor, such as sky view, fish-eye correction, and guideline indication. ● *Figure 18.9* shows these applications. Moreover, as the in-car navigation system gets the steering-angle information from vehicle side, the prediction guideline can be shown on the navigation display. With the spread of vehicle camera modules that can be provided at a moderate price, and the increasing interest in safety, front camera applications are also beginning to become popular. This front camera application is used when vehicle enters the intersection or starts moving. By recording the location when front camera function is used, it is possible to automatically switch over the camera application without any driver handling at the location where the front camera was used.

It has become possible to display the synthesized image from a number of in-car cameras using advanced image processing techniques to eliminate blind spots. This process is executed by a special unit, but this process would be executed by in-car navigation system, using its more powerful processor with multi-core CPUs and graphic processors, if these cameras are connected to car navigation system using vehicle network. The in-car navigation system would display not only the image captured by vehicle cameras but also would be able to display the image processing result that is analyzed by image data processing block using a camera to view front facing from the car. Applications such as object recognition, that is lane recognition, vehicle recognition, traffic light recognition, road sign recognition, and so on, has been developed, but camera application will become increasingly closely tied to car navigation function to



■ Fig. 18.9

Example of back camera application image

complement object recognition and information using map-related data. Displaying arrowed line guidance and highlighting landmark to take a turn, it might enable the driver to recognize the turn point easily by overlaying a live-action image on in-car navigation display. AR (Argument Reality) technology has been advancing, and this technology could be adapted to in-car navigation system so that a facility name attached the facility in a live-action image would be displayed on in-car navigation display. **▶ Figure 18.10** presents an example of the application of object recognition.

The technology of camera application would be advanced chiefly by mainly automotive manufacture striving for the realization of a safer car. The combination of millimeter-wave radar devices and camera devices, for assistance with low visibility in bad weather conditions or backlit conditions, would be an important technology to make precise measurement of cars or other objects in front. In-car navigation system would evolve to become equipment that has “electric eyes” to provide information on surrounding car environment and to inform the driver of danger and risks with a combinations of sounds and pictures.

4.7 Green Technology Application

Global warming has recently become an international concern. The proportion of CO₂ emission caused by automotive is not large at all. Automotive companies aim to reduce CO₂ emission and promote technological developments. Using an in-car navigation



■ Fig. 18.10
Example of the application of object recognition

system that guides to the destination safely and with certainty enables the reduction of CO₂ emission due to driving by avoiding traffic jams and road hazards. In-car navigation systems are one of the successful equipment in helping to reduce CO₂ emission, and these functions continue to be developed. The route planning that an in-car navigation system plans from the car position to the destination is calculated by considering the factors the user selects, which are distance priority, avoiding toll priority, toll standard, or main road priority. Moreover, in-car navigation system that can obtain traffic information can propose a route plan that can be traveled in the shortest time to the destination by avoiding traffic jams caused by heavy inbound traffic. Recently, in-car navigation systems have been developed which also calculate the fuel consumption for each route and propose the route with minimal fuel consumption. The principle is that, as the route is planned using road link which are segments of road network, calculating the average speed for each road link based on distance and time, it is possible to estimate the fuel consumption of a route by calculating the fuel consumption of each section. An application has also been developed which, by calculating acceleration, deceleration, speed, idling time using the information from sensors which the in-car navigation system has, the car navigation detects long idling time, abrupt acceleration and deceleration, and shows the driver a driving evaluation result, shows whether or not they drive in an environmentally friendly way. An in-car navigation system that can provide information on fuel consumption has been developed and launched in the market. It is possible for OEM in-car navigation systems to calculate fuel economy by measuring fuel consumption every second using the precise information on amount of fuel use from the ECU, but it is hard for after market in-car navigation systems to obtain this fuel use information. Generally, fuel consumption is mainly determined by vehicle weight, air resistance in running, rolling resistance of tires. Making use of this principle, it is possible to estimate fuel consumption without connect the in-car navigation to the ECU. This enables the user to confirm the situation of fuel consumption on the in-car navigation display by the

various types of parameters, such as average, instantaneous, and trend fuel consumption, and so on. Sending these data to a fuel consumption ranking site created on the Internet, and comparing them with other drivers raises awareness of the environmental issue. Moreover, uploading the information on fuel consumption condition to the server from in-car navigation system, the accumulated running histories would be analyzed on the server, and it will be possible to construct a road network that reports fuel consumption for every time period in the week. This fuel consumption road network data will be returned to the in-car navigation system to enable it to propose a route plan which will require less fuel consumption. Such in-car navigation system with ecological functions and providing a sense of participation in a “green” environment are expected to continue to be valuable equipment in the future.

4.8 Electric Vehicle Navigation System

Amid increasing awareness of environmental issues, and tax credits for purchasing hybrid or other eco-friendly vehicles, car manufacturers expect the next generation of eco-friendly vehicles to become very popular. In-car navigation system with eco-friendly functions that can display average gas mileage or provide the most fuel-efficient route plan have been developed and brought to the market. In the case of electric vehicles, in-car navigation systems suitable for electric vehicle are required. Electric vehicles have several advantages, including zero-emissions, acceleration performance due to max torque available from start-up, a quiet interior, and low noise impact on the environment, but electric vehicles also have the drawbacks of low range, long charge time, and the inability to notice an electric vehicle that comes near because of its silence. An in-car navigation system that provides appropriate functions which respond to the requirements of electric vehicle is needed to eliminate anxiety about running distance by showing the battery charge remaining or the location of battery charge stations. The aids to eco-friendly driving are “Route planning of most eco-friendly route” and “Evaluation of eco-friendly driving,” which technology solves the two side issues about economic efficiency and environment. Eco-friendly route planning is that the in-car navigation system recommends not the fastest expected arrival time for the destination, but the most electrically efficient route. The rough planning of the eco-friendly route about an electric vehicle works in the same way as for a gas vehicle, but there are two highly important differences: there are no fuel cost while idling and the possibility of energy regeneration while decelerating. These characteristic are similar to hybrid vehicles using an electric motor in combination, but although the single energy source of an electric vehicle makes it easy to manage energy, there is demerit of not being able to use the energy that an electric vehicle could get from other energy sources. The evaluation of eco-friendly driving is mainly used as a means to provide information that will make the driver more aware about an eco-friendly life. This application is a visualization of real-time energy balance or comparison of power-consumption at particular sections. Electric vehicles have the benefit of accurate control of energy when using the electric motor as a drive system. Using the efficiency map that

represents the relationship between revolution speed and motor torque, it allows power consumption to improve by traveling in the most efficient energy condition possible. Unless automated cruising becomes a reality, a highly developed Human Machine Interface (HMI) to assist eco-friendly driving will be required. The visualization of real-time energy efficiency and driving assistant will become more important.

As people who drive electric vehicles have worries about running out of charge, they have particularly large expectations for in-car navigation system to eliminate anxieties about how far the electric vehicle can run and where the nearest electric charge station is. It is said that the relationship between the range of electric vehicle and battery charge is hard to determine because power consumption of instruments such as air conditioner, wiper, headlight, and so on, which are used in response to external environment changes, cannot be ignored. Accordingly, it has become much more important to establish technology that can indicate the range of possible travel on the in-car navigation map and to recommend the most energy-efficient route by deducing the power consumption. It will become more important to provide the driver with accurate information that are location, equipment, opening hour, and waiting time of charge station, to relieve their anxiety about running out of charge. There is also need to allow for the location of charge stations along the route when a route is calculated to a destination. An in-car navigation system that is designed to correspond with electric vehicle would need to have a communication module to get the newest information about charge stations or to check the charge condition that the owner of electric vehicle can confirm remotely using cellular phone or PC. The reason for checking charge status is that it takes long time to recharge a battery, and electric vehicle owner would like to observe the charge condition remotely while the car recharges. If an electric vehicle is recharged at home, the car owner would like to keep an eye on the battery condition using a communication module or a recharge cable that is connected to the in-car navigation system, and in-car navigation system would confirm the recharge function and other vehicle condition status.

Generally, it is said that vehicles spend more time parked than running. It seems very possible that charge cable is plugged into the vehicle while recharge in cases where the electric vehicle is at home. And it may be that an electric vehicle is viewed as one of home's electric appliances. The same holds for in-car navigation system. If the in-car navigation system is always on, it can handle various new applications such as, for example, fixed-point observation network using sensors of electric vehicle, or content storage. Taking advantage of the silence in an electric vehicle, the in-car navigation system could provide new AV functions and ways to make things enjoyable and exciting that are suitable to electric vehicles.

One of drastic change in life style that electric vehicle creates is to recharge at home. As the electric vehicle has a large-capacity battery, this vehicle battery can be also used as home battery. The electric vehicle battery can help to smooth the electric power that is generated by energy of nature, such as solar and wind power, at home. In such situation, the in-car navigation system would be used to control discharge and charge operation efficiently by predicting the electric generation possible from renewable energy using weather information. This control of discharge and charge would enable it to buy electric

power at low prices or to sell it at high prices. It would be better to make the in-car navigation system include an application for a dynamic pricing system for electricity based on demand and supply.

The evolutionary process of in-car navigation system on electric vehicle may be starkly different from technological advances of conventional in-car navigation systems.

5 Conclusion

Tracking back the history of in-car navigation systems and seeing the in-car navigation function, we see that the in-car navigation system has advanced from operating as a stand-alone, to being an integrated system that contains various equipments and has enhanced in-car navigation functions to connect to the outside world environment. Furthermore, the transition of storage media for map-related information, from CD-ROM, DVD-ROM, to HDD, and flash memory, has had a great impact on capability of in-car navigation systems. The display of an in-car navigation system that has gained big display size and high resolution has enhanced the level of its expressiveness, acquiring the capabilities of a graphic processor that has evolved due to advances in SoC technology.

When the in-car navigation system gained the function of supporting the driver by providing road traffic information, the in-car navigation system stepped into the second stage, no longer working as stand-alone systems. Current in-car navigation systems that have the function of interactive communication to connect a cellular phone or communication module, not only getting the traffic information but also uploading prove information, have stepped into third stage. Right now, in-car navigation systems have entered a new territory to provide useful information using the network. This means that in-car navigation systems have enhanced the function of information acquisition from various Internet servers at will so that the functions and data need not be stored in in-car navigation system.

An in-car navigation system with audio and video functions has become an established normal style, but the AV function also has been becoming vastly more important to link portable music devices or smart phones, allowing access to functions the in-car navigation does not have. Thus an in-car navigation system would enhance Internet connectivity, allowing the user to access various music and video content.

Due to international concern about global warming, green technology has become important in any in-car navigation system. Especially using prove car system technology, an in-car navigation system would provide a better route plan to further reduce fuel consumption. In-car navigation systems for electric vehicle have started attracting attention to provide proper route planning within the present charge range, or information on the availability charge stations using the communication function. More attractive applications and services would be expected.

Having a large display that has high ability of expression, the in-car navigation display that takes center stage and has been playing an important role to display and operate the various information a driver needs. In-car navigation systems will enhance the functions of navigation, entertainment, and information, and will be a must-have feature in vehicles to come.

References

Fujimoto H (2001) World wide vehicle navigation system using KIWI format. Denso Tech Rev 6(1):29–34

Tagami K et al (1983) “Electro Gyro-Cator” new inertial navigation system for use in automobiles.

SAE Technical Paper 830659, SAE International, New York

VICS (2006) The history of VICS. Vehicle Information and Communication System Center, Japan

Section 5

Driver Assistance

Azim Eskandarian

19 Fundamentals of Driver Assistance

Azim Eskandarian

Center for Intelligent Systems Research, The George Washington University, Washington, DC, USA

1	<i>Introduction</i>	493
1.1	Background	493
1.2	Chapter Coverage	495
1.3	Driving Tasks	495
2	<i>Driver Perception–Response</i>	498
2.1	A General Human Response Model	498
2.2	Human Reaction Time	499
3	<i>Driving Tasks and Other Influences on Perception–Response</i>	504
3.1	Cognitive Perspective of Driving Tasks	504
3.2	Driver Distraction	507
4	<i>Driver Assistance Types and Levels</i>	510
4.1	Classification According to Level of Intervention	512
4.1.1	Informational	512
4.1.2	Warning	513
4.1.3	Partial (Semi) Control	514
4.1.4	Automatic Control	514
4.1.5	Autonomous Control	515
4.2	Classification According to Adaptability	515
4.3	Other (Temporal) Classification of ADAS	516
5	<i>Integrated Safety</i>	519
6	<i>Human–Vehicle Interface</i>	520
6.1	Feedback to Driver	520
6.1.1	Age Consideration	522
6.1.2	Mental Workload	522
6.1.3	Vigilance	523
6.2	Receive Driver’s Input	524
6.3	Actuations: Execute Automatic and Autonomous Action	524

7 *Testing and Evaluation*525

8 *Samples of Driver Assistance Systems (Case Studies) and Future
Research Needs*526

8.1 Examples of Existing ADAS 526

8.1.1 Brake Assist System Supports Braking by Intervention 527

8.1.2 Adaptive Headlights Enhances Visibility and Safety at Night Driving 527

8.1.3 Blind Spot Detection Augments Driver’s Field of View and Provides
Visual Feedback 528

8.1.4 Lane Departure Warning Improves Situational Awareness, Warns/Alerts
driver/Lane Departure Control Improves Situational Awareness,
Augments Steering Control 528

8.1.5 Driver Assistance Pack 530


8.2 Future Research 530


9 *Concluding Remarks*531

Abstract: Driving is a complex task of strategic decision making, maneuvering and controlling the vehicle while responding to external stimuli, traffic laws, and imminent hazards. Driver's cognitive perception and reaction, physiological, and psychological capabilities along with experience, age, and many other factors play a major role in shaping the driving behavior and the skills to control the vehicle. Driver assistance systems are designed to support the driver in performing the primary driving tasks and the secondary in-vehicle tasks that may be required (operating radio, etc.). The goal of driver assistance or ADAS (advanced driver assistance systems) is to enhance safety, comfort, and efficiency of driving by intervening in the handling aspects of the vehicle and supporting the secondary tasks for comfort, navigation, etc. Driver assistance deals with the environment in terms of sensing and responding, the vehicle in terms of sensing and actuating electromechanical systems, and most importantly the driver in terms of augmenting information, enhancing sensing capabilities, and assisting in control functions. This chapter examines various aspects of driver assistance system including driver cognitive perception–response, system types and classifications, integrated safety, man–machine interface, and evaluation of effectiveness. This chapter concludes with listing existing ADAS and research needs.

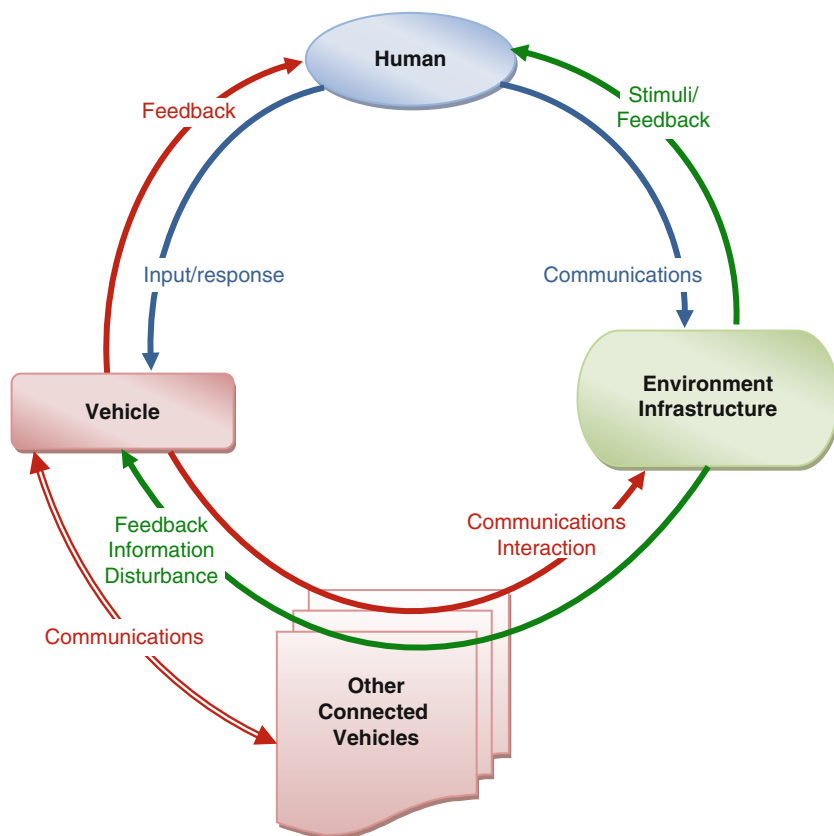
1 Introduction

1.1 Background

Driving is a learned control task performed by a driver. A human driver develops the necessary skills to control a car through training. In the most simplified terms, relevant to the vehicle, these skills involve starting and accelerating a car to a desired speed, braking to slow down or stop the vehicle, steering the vehicle on a desired trajectory, i.e., moving straight, keeping within a lane, changing a lane, and negotiating curves and turns as necessary. The most basic human (driver) control functions or tasks, which actuate a vehicle, are accelerating, maintaining speed, braking, and steering. In performing these functions, the driver also interacts and responds to the environment, which includes both the internal in-vehicle and the external environment (roadway, signals, markings, other traffic, hazards, etc.). This fundamental interaction of Human–Vehicle–Environment is depicted schematically in  [Fig. 19.1](#).

The interrelationships illustrated in  [Fig. 19.1](#) indicate that the interaction among all three elements needs to be considered thoroughly for the design and development of any vehicular safety or comfort systems.

In addition to the basic vehicle control functions, a major part of the driving task is to interact and respond to the surrounding environment and traffic. These include obeying the traffic laws, performing the various maneuvers, and responding to dynamic emerging situations. Examples in a normal driving condition are adjusting the speed, stopping, changing lanes, passing, merging in and out of highways, maneuvering around obstacles,



■ Fig. 19.1
Schematic of driver-vehicle-environment interactions

and many other responses to the impending traffic situations. As the basic driving skills develop, drivers also need to master responding to unpredictable situations, i.e., conditions that arise beyond normal driving circumstances. These could be natural hazards such as falling rocks, slippery icy road surface, animal crossing or man-made such as sideswiping cars, red-light running cars, and bicyclist or pedestrian infringement, etc.

As drivers learn to drive, they develop the necessary skills to both control the vehicle and respond to the changing environment. However, the skill levels differ significantly among drivers. Educational, cultural, social, training, and law enforcement factors all influence drivers' behavior. All of these factors combined with the physiological and cognitive abilities of the driver affect how a driver controls a vehicle. They constitute what is known as the "driving behavior" which includes driving decision making, responses, and control functions. Therefore, with all these influencing factors, understanding and characterizing one's driving capabilities and behavior can be very complicated.

Driver assistance systems refer to systems that help the driver in various tasks and functions; their overarching purpose is to improve driving safety, comfort, and efficiency. In order to develop support systems, or assistance for drivers, the general driving behavior and all the associated factors that affect that behavior and performance must be considered.

1.2 Chapter Coverage

This chapter covers fundamental issues in understanding and developing driver assistance systems. This is neither intended as a complete guide to develop new systems nor a full listing of all available systems. The goal of this chapter is to cover major issues, challenges, and expert views in development of driver assistance systems with key references to enable the readers to further pursue their specific areas of interest. A full coverage of this subject is indeed the goal of this entire handbook and many aspects of driver assistance are covered in other sections of the handbook as well.

Introductory parts 1.1 and 1.3 of this chapter cover the basics of driving tasks. Part 2 covers human perception–response capabilities, which helps define how drivers perceive and react to stimuli during driving, and hence develop an understanding on how to support these functions. Part 3 includes additional influences on driver perception–response. Part 4 describes different methods of classification of driver assistance systems, e.g., according to their level of intervention with the driver. These classifications characterize the requirements of the system to be designed. Part 5 briefly touches upon the role of driver assistance in a holistic view of the integrated vehicle safety. Part 6 further examines the human–machine interface aspects of driving. Part 7 discusses the reliability of driver assistance systems and explains standard testing and evaluation process. Part 8 presents case studies of existing and prototype advanced driver assistance systems (ADAS) and some future research needs. Finally, Part 9 summarizes the concluding remarks.

1.3 Driving Tasks

Michon defines three operational levels for modeling driving, which include the cognitive process (Michon 1985). These are strategic (planning/decision making), maneuvering (tactical), and control (operational) levels. Most experiments tend to determine or model specific control and maneuvering functions because empirical determination of cognitive capabilities (higher levels) is very difficult. The variability of the driving tasks and combinatory effect of external stimuli affecting drivers' decision making adds to the complexity and difficulty of developing driver cognitive models. Holistic research at all levels is still required to arrive at verifiable and validated models of driving behavior.

A long list of driving tasks and subtasks are identified by McKnight and Adam's early work (McKnight and Adams 1970a, b) for driver education and training purposes. Their analysis classifies the driver actions to 43 main tasks that are further divided into 1,700 subtasks, each

further divided into more detailed activities. This detailed classification of driving tasks, although useful for training and other investigations, does not necessarily provide a human driving behavior and performance model from a cognitive–control perspective (Godthelp et al. 1993). Task descriptions, which can be expressed in terms of perception–response behavioral models or other methods that substantiate driver’s workload demand (analytically or empirically) are required to aid in development of support systems.

The GIDS (Generic Intelligent Driver Support) program provided behavioral corroboration on the task descriptions and basic reference models, which describe the optimum driver action and vehicle motion characteristics for particular maneuvers (Michon 1993, p.24). A thorough analysis of driving tasks in this context (i.e., similar to GIDS), along with consideration of accident data, technological development in vehicle dynamics controls, and human–machine (human–vehicle) interface (HMI) can help identify what aspects of driving can indeed be supported in the form of a driver assistance system. The goal of such system would be to reduce the driver workload, increase safety and efficiency, and enhance driving comfort.

All elements of vehicle feedback to the driver are also sources of mental and manual workload. The in-vehicle controls (steering wheel, brake and gas pedals, etc.) and instrument panel provide the necessary feedback to the driver, part of which is used for the driving function and others for information, comfort, trip planning, and entertainment, etc. For example, the steering system and suspension are part of vehicle design which gives a direct feedback to the driver for vehicle control functions: They affect how a driver steers to negotiate a curve, avoid an obstacle, or in general handle a vehicle. On the other hand, mirrors provide enhanced field of view of the surrounding, the speedometer provides a visual feedback of the vehicle’s speed, or a navigation system provides information and step-by-step directions for trip planning. A driver’s glance at the mirror, speedometer, or navigation system creates a visual demand.

Therefore, in addition to the main driving tasks that are cognitively, visually, and manually demanding, there are other in-vehicle tasks that create workload and demand driver’s attention. These are secondary tasks that the driver performs and can be separated into five categories based on the resources demanded of driver (Wierwille 1993). Wierwille classifies the five categories as:

1. Manual Only:
Done by drivers hand or foot without needed visual reference after sufficient initial practice or memorization of the function; example: sounding horn, adjusting low/high beam
2. Manual Primarily:
Driver glances (uses vision) to find control and then performs the rest of the task manually; example: turning on radio
3. Visual Only:
Tasks that have no manual input and are visual only for basic checking or more complex information gathering; examples: reading speedometer, or viewing navigation system

4. Visual Primarily:

Rely heavily on vision but require some manual input; example: changing operation mode on audio system (time display to radio) or music selection, accessing different features of navigation system (different viewing mode, etc.)

5. Visual–Manual:

Require interactive visual–manual demands in which driver gathers information to make manual input or performs manual tasks to access desired information or visual feedback. These are most demanding of driver resources as driver performs adjustments visually and manually between displays and controls which can take the driver's attention away from driving; example: manually searching on radio frequencies; operating cell phones

Although in-vehicle information systems are intended to enable safe and comfortable driving, their counter intuitive design could cause additional demand or workload for the driver.

In the context of driver assistance, the in-vehicle systems provide the information and functionality needed to operate a vehicle. While the design variety and capabilities vary from vehicle to vehicle, the basic functions are similar. The notion of *driver assistance* is therefore not a new idea, but with the emergence of new technologies, the vehicle's support system for the driver can be further enhanced. This may have many facets. The sensing capabilities can be enhanced by the use of radars and vision system. The control abilities can be augmented by completing a function which the driver starts but cannot perform optimally, e.g., assisting in the braking function. The overall driving experience can also be improved by automatic functions, which are transparent to the driver, like the electronic stability control or traction control that sense the situation dynamically and are actuated to automatically enhance vehicle trajectory control. It is in this context of new technologies that the term “driver assistance” is consummating.

Driver assistance interacts with the *environment* in terms of sensing, the *vehicle* in terms of sensing and actuating electromechanical systems, and most importantly the *driver* in terms of augmenting information, enhancing sensing capabilities, and assisting in control functions.

The term “advanced driver assistance systems,” commonly known as ADAS, refers to systems which improve the vehicle safety, comfort, and efficiency by intervening in the handling aspects of the vehicle.

Earlier, ADVISERS project (Stevens and Parks 2001), a government and industry funded European project identified six areas of development of ADAS that would have high potential of impact and likelihood of implementation:

1. Variable speed limiter (intelligent speed adaptation system – ISA)
2. Adaptive cruise control (ACC)
3. Platooning (Interurban)
4. Driver monitoring
5. Road surface monitoring
6. Lane keeping

Of course, by now some of the above systems have already emerged in the market, as explained in the final part of this chapter. For example, adaptive cruise control (ACC) and lane departure warnings (LDWs) are available by several manufacturers now (ACC uses radars to maintain cruise control even at presence of other vehicles in the same lane; LDW alerts when drivers are departing a lane unintentionally). Also, informational ISA (intelligent speed adaptation systems, which provide speed advisory feedback to drivers) are now available on navigation systems; they warn the driver when the posted speed limit or a set speed threshold is exceeded by the car. Other features have been the subject of R&D and prototype development. For example, steering control for evasive maneuvers has been demonstrated by various manufacturers. Full autonomy (driverless cars) including navigation has also been developed in DARPA (US Defense Advanced Research Projects Agency) Grand and Urban Challenge competitions from 2004 to 2007 (Behringer et al. 2004; Ozguner et al. 2007; [Web Site: DARPA](#)). DARPA Grand Challenge demonstrated driverless vehicles finishing a long off-road trail with challenging terrain and the DARPA Urban Challenge demonstrated autonomous driving in an emulated city environment obeying traffic rules and signs. Broggi et al. (2010; Broggi 2011) further demonstrated the challenges of a long haul 3 months autonomous driving through an intercontinental trip, which took place during the 2010 World Expo in Shanghai, China.


While driverless or autonomous vehicle removes the driver from the task of driving in the future, the current reality of driver assistance systems is that they still need to work in conjunction with the driver and support driving functions as seamlessly and unobtrusively as possible. Therefore, the most important thing in the development of driver assistance systems is how to relieve driver's workload and support the driving function flawlessly. To this end, it is imperative to understand the human aspects of driving.

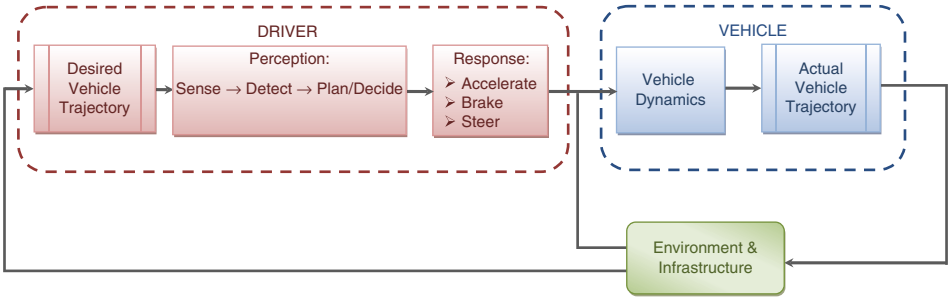
2 Driver Perception–Response

2.1 A General Human Response Model

The “driver” aspects of the driver assistance require a more fundamental understanding of the driving functions from a human cognitive and perception–response perspective. Basically, the driving, similar to other human motor skills requires the following steps:

1. Sensing
2. Perceiving
3. Planning/Deciding
4. Responding/Action

The above happens in a continuous feedback fashion during driving. Alternate definitions for these are also seen in the literature. For example, the steps of Detection, Identification, Decision, and Response are alternate descriptors for the same functions (Olson et al. 2010). The schematic of these functions are depicted in  [Fig. 19.2](#).



■ Fig. 19.2
A simplified feedback cycle of a driver's functions

Although this breakdown may be suitable for detailed analytical study of the physiological and cognitive behavior of the driver, in reality, these functions cannot necessarily be measured separately, and thus are better represented by two main functions of perception and response. The perception includes the first two functions (detection and identification) and response includes the latter two functions, i.e., the planning/deciding and the actual motor action. Alternately, planning and deciding may also be considered as part of perception, as depicted in the figure, and response include only the actual motor function. The main reason for this analysis is that for performance evaluation purposes or driver reaction measurements, it is essential to know how long it takes for a driver to react to an external stimulus. For example, safety engineers need to know how long it takes to initiate braking after a driver sees a stop sign (in normal driving) or an incoming obstacle (in emergency situations). Therefore, in experimentation of driver responses, typically a well-defined external stimulus is presented and the drivers' response is observed for a specific scenario.

The most important purpose of this exercise is to find the time it takes for the driver to respond to an external stimulus. This is called the perception–response time (PRT) and can be measured with some level of success. The perception–response time is the critical design elements for many vehicular safety systems, which interact with the driver. Therefore, any driver assistance system should take this important factor into consideration. However, due to the extreme variability of the drivers' abilities and the variety and complexity of the situations that may arise, it is very difficult to determine the perception–response time. Even for one driver, the fatigue condition, the day or night, or a variety of environmental and traffic conditions could cause a difference in perception–response time.

2.2 Human Reaction Time

The most commonly known law dealing with reaction time is the Hick's Law of 1952, which determines the reaction time depending on the number of alternatives or possible choices presented to a person (Hick 1952). It relates the logarithm of the number of

choices (presented alternatives) with the mean reaction time for a correct identification. In Hick's law, the average reaction time T required to choose among n equally probable choices is given by:

$$T = b \cdot \log_2(n + 1) \quad (19.1)$$

where b is an empirical constant. This model predicts the reaction time to external stimulus choices or information bits increases logarithmically. Hyman (1953) varied the information conveyed by a stimulus in three different ways and showed the reaction time to the amount of information in the stimulus obeyed a linear regression for each of the three ways. The three ways were (1) the number of equally probable alternatives (Hick's law), (2) the proportion of times it could occur relative to the other possible alternatives, and (3) the probability of its occurrence as a function of the immediately preceding stimulus presentation.

Although Hick's law does not directly provide a reaction time, many researchers estimate 0.10–0.15 s as the time to take one action (of one choice). Since the variations in tasks and human abilities and conditions are so large, it is difficult to quote one number as the reaction or response time to a stimulus. Time to respond to a stimulus varies greatly among different tasks and even within the same task under different conditions; this time can range from 0.15 s to many seconds and is highly variable (Green 2009a). Although 1.5 s is used for PRT (perception–response time) by many experts, in some cases, even the concept of perception–reaction time (for driving or other situations) as presented here (and in the literature) may not apply. Therefore, all data must be interpreted only within the scope and the conditions of the experiments whether driving or others.

Transportation Research Circular 419 (Hall and Dudek 1994) reported drivers' perception–reaction times in emergencies based on a Canadian experiments conducted on 40 subjects with equal age distribution below and above 35 years (Wilson et al. 1989). Subjects driving at 60 Km/h were presented with obstacles 40 m away; they were instructed to brake to stop the vehicle in one experiment and steer in another. Brake reaction time and steering reaction time from their experiments are shown in [Table 19.1](#). The results show the design criteria of 2.5 s for braking and steering is a conservative choice. It must be noted that in this study subjects were alerted and practiced the experiment several times. Tapani (2009) provides a detailed literature review of various human reaction and response times but in the context of learning the delays related to closed-loop controls in which human is involved. Among many findings in this reference, it is interesting to note a few related to the driver assistance systems, namely auditory, visual, and touch sensing that take progressively longer time to reach the brain, with 8–10 ms, 20–40 ms, and 155 ms, respectively.

Green (2000, 2009b) shows three sets of values for braking response time depending whether the driver is informed about the test. Green shows the reaction time from detection (of obstacle) to moving foot from gas to brake pedal are 0.7–0.75 s, 1.25 s, 1.5 s for expected, unexpected, and surprised drivers, respectively. In each case, the driver will have different perceptions in an emergency maneuver. In the first case (expected driver), the best estimate is 0.7 s; 0.5 s for perception and 0.2 s for the movement in normal

■ Table 19.1

Brake and steering perception–reaction times in emergencies involving obstruction in roadway (Data duplicated from Hall and Dudek 1994)

Brake perception–reaction time (s)									
Driver perception time			Driver reaction time			Vehicle braking response time		Total	
Mean	Std. Dev.	99th percentile	Mean	Std. Dev.	99th percentile	Mean	Std. Dev.	Mean	99th percentile
0.56	0.16	0.90	0.28	0.10	0.58	0.12	0.03	0.96	1.6
Steering perception–reaction time (s)									
Driver perception–reaction time							Time until vehicle abreast of target		
Mean			99th percentile						
1.00			1.8				2.4		

*99th percentile is found by fitting the data to a normal distribution

traffic conditions. In the second case (unexpected driver), the time to perception is increased to a second for a total of 1.25, which is valid when the driver must respond to traffic signs, etc. The latter case is the situation where a pedestrian or an unexpected obstacle appears in front of the car. The reaction time takes into account the distance of the obstacle and the position of the obstacle relative to the car (in the field of view). If it is in the driver's peripheral vision it takes 1.2–1.3 s perception time for the movement. If the obstacle appears in front of the driver, the total time may be a few tenths of second faster. Also when driving at night the contrast decreases and therefore the reaction time increases from 0.20 to 0.25 s.

Card et al. (1983) divide human information processing model into three time steps: (1) perception time (T_p), (2) cognition time (T_c), and (3) motor time (T_m). The first two, as indicated above, are grouped together as one possible measurable step designated as perception–cognition time ($T_{pc} = T_p + T_c$). Hidetoshi, et al. use this method in a driving simulator experiment to generate probabilistic models on driver perception–response time among other things (Nakayasu et al. 2010). They have conducted a driving simulator study with 37 male subjects (26 and 11 men aged 21–24 and 60–64 years old, respectively) to evaluate the perception, perception–cognition, and motor time during braking (among other things) for three groups of trained, untrained, and aged drivers. In their braking experiment, they divide the braking into two segments of T_p , time for perception, and T_m , time for motor action, e.g., for the execution of the brake by the driver (this T_p means the same T_{pc} from the first experiment). During perception time the vehicle is still moving at a constant speed and during braking time, T_m , the vehicle speed reduces linearly to zero until the stopping point. The braking was done at 80 km/h on the onset of observing a random truck and the drivers were to complete the braking action immediately to

■ Table 19.2

Estimated values of response time at 80 Km/h in driving simulator (Data taken from (Nakayasu et al. 2010))

	T_p (s)	T_{pc} (s)	T_m (s)
Trained subjects			
Median	0.473	0.586	2.404
Mean	0.481	0.599	2.663
Untrained subjects			
Median	0.512	0.729	2.440
Mean	0.523	0.717	2.379
Aged subjects			
Median	0.616	0.870	2.547
Mean	0.599	0.869	2.504

achieve the shortest stopping distance. Fifteen iterations were performed for each braking. The T_{pc} was calculated with another set of visual simulator experiments in which a randomly appearing truck was introduced requiring the reaction of the driver. In this experiment, two visual stimuli appeared as “target” sample requiring a human response and “standard” sample to which a response was not needed. The target and standard stimuli were introduced 50% of the time each to ensure subjects do not learn the habitual behavior of always responding to one stimulus. ▶ Table 19.2 shows excerpts from their experimental data in a different format, where means were derived from the fitting of the data to different probability distributions (e.g., Weibull and others) (Nakayasu et al. 2010).

In addition to the above, (Nakayasu et al. 2010) propose a HCR (human cognitive reliability) probability-based model with performance shaping factors including training (skill) with various control parameters and conclude that there are differences between the cognition process of the trained versus untrained drivers. They conclude from their experiments that driver behavior for T_{pc} indeed has both perception and cognition processes but T_p alone does not have cognition process. They also conclude the driver experience is one of the most important factors for the reaction time. This factor significantly reduces the risk when an obstacle appears in front of a motorist.

In another simulator study, Mahmood and Easa evaluated the reaction time of drivers in a car-following environment (Mehmood and Easa 2009). A total of 60 subjects (32 male and 28 female) aged from 18 to 70 years were tested on three different scenarios in which a car appeared before them but with a different rate of appearance in each scenario. The objective was to evaluate the reaction time of drivers. Two different reaction times were defined as brake reaction time (BRT: when the lead vehicle is braking and its brake light is on) and acceleration/deceleration reaction time (ADRT: when the driver reacts to adjust his/her speed using the gas pedal only). Three kinematic conditions introduced as normal, surprised, and stationary provided the different levels of urgency and expectancy based on

the braking behavior of the lead vehicle at different speeds and spacing. The BRT and ADRT were the dependent variables and driver's age, gender, driving experience, driving intensity (driving hours per week), vehicle speed, and spacing were independent variables. Their results showed there was a significant difference in the BRT at normal, surprised, and stationary scenarios and driver's age, gender, speed, and spacing were found to be significant variables for the BRT in all scenarios. The results also showed the BRT increases with age and that driver's age and gender were significant variables for the ADRT. Their results also showed a brake reaction time range of 0.892–1.934, 0.632–1.004, and 0.558–0.937 s for normal, surprised, and stationary scenarios, respectively, as a function of various speed/distance levels ranging from 60/10 to 100/40.

It should be mentioned that there are differences in simulator versus real-life reaction times. Drivers' control of vehicle could be significantly distorted with simulator latency. However, with improved computer speeds and better graphics and simulator designs the latency problems are reduced significantly. In both stationary and mobile platform driving simulators, creating true braking inertia effects is a challenge. Low to medium fidelity driving simulators lack the realism of braking deceleration or imposing jerk on driver. This is clearly evident in stationary simulators at which only visual and audio queues exist for braking feedback. The secure feeling in a simulator may also cause a higher level of risk taking and hence skewing the results of any study involving emergency maneuvers.


As mentioned earlier and commonly known by many experts, there is really no average time for the human reaction, although 0.7s is commonly considered as the optimal time of reaction. Reaction time is highly dependent on the task and all other conditions of the stimulus and the driver. Olson and Sivak (1986) experimentation of 64 drivers, with ages 18–64, found 95% of drivers in both the young and old age groups showed a PR (perception–reaction) time of 1.6 s when drivers were confronted with an unexpected roadway hazard (obstacle) with subsequently cresting a hill. This suggests the standard 2.5 s used for stopping sight distance is sufficient for all age groups. Depending on the situation and driver's abilities, the overall perception–response time can be anywhere from 0.5 to 3.5 s or more (Olson et al. 2010, pp. 377.)

Olson et.al. (2010, pp. 375–488) present a detailed review and analysis of driver perception–response time and many factors affecting it, as well as many other driver issues, including topics in weather, older driver, and distractions.


With all variables being equal, the reaction time still varies among drivers. The review of reaction time here has focused mostly on braking or car-following cases because braking and speed reduction is one of the most important safety factors. Also braking is among better defined functions with suitably distinct control variables. Determination of perception–response time (PRT) for other driving tasks are even more complex with numerous other control variables influencing the PRT value. Development of ADAS for any driving conditions or task must incorporate a careful examination and study of the respective driver perception–response issues as delineated in the above examples but with a cautionary note that these models may not apply in all situations, especially conditions in which human cognitive and response abilities are drastically affected by hazardous stimulus, shock, fear, etc. Some of these are described in the following section.


3 Driving Tasks and Other Influences on Perception–Response

3.1 Cognitive Perspective of Driving Tasks

During normal driving, a driver performs multiple tasks continuously and concurrently, e.g., maintains the vehicle speed by adjusting the pressure on the accelerator or brake pedal and keeps the vehicle at the center of a lane by corrective steering. Each of these tasks involves all the steps illustrated in  Fig. 19.2.

Consider a simple driving task like maintaining a desired vehicle speed in a straight lane. The driver sensing is visual, audio, and haptic during driving. The driver senses the speed of the vehicle by viewing the surrounding or the speedometer, feeling the sound, feeling the vibration and physical inertia, or a combination of these senses. Then, he/she perceives the speed and has a notion of safe or desired speed, and decides to either maintain or change to a desired speed. Finally, the driver adjusts the vehicle speed by accelerating or braking. Therefore, in the context of driver's function, many driving tasks during normal driving can be analyzed through this cycle of Sensing→ Perception→ Planning/Deciding → Responding/Control Action.

Enumeration of all driving tasks is beyond the scope of this chapter. However, for the purpose of illustration, a few selected driving tasks are analyzed in the same manner described above with respect to the four fundamental processes and the two measurable processes of perception and response.  Table 19.3 shows the details of the selected driving tasks as analyzed according to the human perception and response sequences.

Examples in  Table 19.3 are standard vehicles control functions in traffic under normal circumstances without disturbances. The subtasks and the four cognitive/control functions can be influenced by traffic, surface conditions, or other environmental disturbances. For example, during a lane change, a large pothole or an icy spot on road surface may be encountered to require additional sensing/perception and decision/control reactions. This could lead to a delayed steering to pass the undesirable surface condition. In a more critical scenario, a fast approaching vehicle or motorcyclist may appear on the adjacent lane in vehicle's blind spot, requiring a quicker driver response, e.g., a change of plan or a delayed lane change. In this scenario, crashes could occur if the driver fails in two respects: (1) to see (sense/detect/perceive) presence of other vehicles or objects in the adjacent lane, or (2) see but fail to respond (decide/control action) by a timely refraining from steering or a delayed execution of lane change. A driver assistance system for this particular situation is now available, which senses the presence of vehicles or objects in the blind spot zone continuously and shows a light (e.g., green or amber) near the side-view mirrors to either allow driver's proceeding with the lane change or warn of existing danger. The blind spot hazard lights must provide continuous monitoring and real-time warning to the driver. Thus, in this example the blind spot detection system helps by enhancing the drivers' sensing and perception (detection) capability, but the decision and the motor action is still left to the driver. This system does not help the mentioned case (2). A higher level of support, may stiffen the steering wheel in the direction of the

■ Table 19.3

Sample driving tasks according to a human cognitive perception–response process

Sample driving tasks	Human processing			
	Sensing →	Perception →	Planning/Deciding →	Responding/Control action
1. Maintaining a desired speed	Senses vehicle speed: visual, sound, haptic (vibration), inertia	Recognizes the speed against a desired value	Decides to press gas pedal to increase, or brake pedal to decrease	Presses the gas pedal or brake pedal with an appropriate force and rate of application
2. Lane keeping	Senses vehicle position relative to centerline of lane: visual, lateral force, or vibration (in case off lane on rumble strips)	Recognizes the vehicle location relative to prior position (off center)	Decides to steer right or left to bring the vehicle to the center	Moves steering wheel to right or left with a corrective action
3. Lane changing	Observes adjacent open lane: visual, direct and in the mirrors, and auditory if other vehicles present	Recognizes the lane is open for merging, or waits until it opens	Two steps: 1. Before sensing, plans to change lane 2. After sensing and recognition of open lane, decides to execute plan or wait	Moves steering wheel to right or left and may adjust speed as needed
4. Reacting to traffic signs, obstacles, and other disturbances (example of reacting to obstacles)	Sees the obstacle: visual	Recognizes the obstacle is a potential hazard that needs to be avoided	Plans to: 1. Brake and/or 2. Steer immediately	Presses the brake pedal and begins steering at a (high) rate commensurate with the perceived distance

potential hazard and hence hinder or prevent the actual lane change. This haptic system provides an additional feedback (or warning) to the driver to be attentive to the situation and helps to partially mitigate the possible collision with the items in the blind spot.

Note that the above analysis is true for most normal driving tasks but not for all situations. During unusual circumstances and hazardous situations, while the same cognitive and control functions are in effect, they may be drastically affected by other human feelings including fear, anxiety, confusion, and shock, all usually resulting in a more severe loss of control. These feelings could have drastic effects on all aspects of

human feedback cycle during driving. They can delay sensing or a hazard could happen so suddenly to prevent a timely sensing all together. Fear or anxiety could incapacitate human's perception and decision ability. Similarly, they can delay a proper reaction or could cause a totally wrong reaction. For example, in avoiding a suddenly emerging obstacle, a driver may press the gas pedal instead of the brake pedal or steer further into the obstacle rather than away from the obstacle. Under these extraneous circumstances, the best assistance alternatives are automatic control of the vehicle if sufficient sensing and actuation capabilities exist. However, it is very difficult to sense and measure these conditions reliably for all possible encounters. The best approach is to have extremely reliable situational awareness, along with vehicle state estimation and trajectory predictions, and enable autonomous control of vehicle that could ensure the safest probable outcome of the hazardous situation. This could be a combination of optimum braking and steering. The societal acceptance and legal implications of such full autonomy is the subject of many discussions, some of which are presented in later chapters of this handbook. Steps toward that goal for well-defined scenarios are already in place. Brake assist system is one example of a partial automatic control of vehicle when an imminent danger is present, but the triggering of this system still depends on the driver's initiation of braking.

As emphasized here, the overall response of the driver is a function of many things including the motor skills, training, physical and cognitive capabilities, reflexes and the ability to handle unusual circumstances, among other human conditions. Therefore, braking, steering, or evasive maneuver data-based models (probabilistic, regression, etc.) developed on the perception-response hypothesis, although important in fundamental understanding of the human-vehicle interactions, should always be used cautiously in development of ADAS. The development of ADAS for such situations is best to rely on how to mitigate the situations automatically as much as possible to provide the safest countermeasures.

Among many variables and conditions that affect the reaction time and driver response are:

- Age
- Gender
- Driving experience (skills)
- Urgency
- Mental load
- Visibility
- Impairment
- Fatigue
- Distraction

These variables are discussed at length in various chapters of *Automotive Ergonomics* edited by Peacock and Karwowski (1993). While each of the above conditions deserves a complete discussion, the more recent challenges of driver distraction are briefly presented here as a major growing problem requiring new ADAS countermeasures.

3.2 Driver Distraction

Distraction can be visual (eyes off the road), manual (hands off the wheel or feet off the pedal during braking), and cognitive (mind off the driving task) (Ascone et al. 2009). The use of cell phone and texting is the most alarming among all distractions because it involves all three types of distraction (Web Site: US DoT 2011). Even in a normal driving environment, e.g., straight road, minimal traffic, etc., clearly, distraction will increase both the perception and response time of the driver and hence reduce drivers readiness in responding to emergency situations. In case of texting, the distraction is often severe and can lead to a loss of vehicle control, even without presence of external hazards. During texting, the driver reaction or available response time is too short to provide corrective action even in normal (nonemergency) cases of lane departures or negotiating curves, etc. The driver ability to formulate (decide) the corrective action could be drastically influenced by a distracting cognitive load alone.

Distraction may also be caused by external factors, billboards, unusual traffic patterns, undesirable lighting, etc. The inattention of drivers and distractions are among the most important causes of accidents. More than 75% of car crashes were found to involve some form of driver inattention; in more than a third of these collisions, cell phone usage was mentioned as the primary form of inattention (Neale et al. 2005).

In 2008, 16% of fatal crashes and 21% of injury crashes were reported to have involved distracted driving according to Fatality Analysis Reporting System (FARS) and the General Estimates System (GES), respectively (Ascone et al. 2009). The same source indicates according to National Motor Vehicle Crash Causation Survey (NMVCCS), approximately 18% of all crashes attributed to the driver involved distraction and during the 100-Car Naturalistic Driving Study, driver involvement in secondary tasks contributed to over 22% percent of all crashes and near-crashes. ADAS dealing with distraction and enhancing safety is, therefore, in great demand. “The proportion of fatalities reportedly associated with driver distraction increased from 10% in 2005 to 16% in 2009” (Web Site: US DoT 2011). The younger drivers at age group under 20 are at higher risk of distracted driving.

The US Department of Transportation (DoT) has started major initiatives in tackling this problem to end the dangerous practice of distracted driving (Web Site: US DoT 2011). While educational and awareness campaigns are underway and law enforcement is active, a large body of driver distraction research has started during the past few years in an attempt to quantify the problem (Strayer et al. 2003; Strayer and Drews 2004, 2006; Drews et al. 2004; Just et al. 2008; Laberge-Nadeau et al. 2003; Dingus et al. 2006; Madden and Lenhart 2009). Research continues to be conducted to characterize the dangers of this growing problem and has spearheaded specific bans on texting while driving in 19 states and the District of Columbia (Web Site: US DoT 2011). To prevent accidents caused by texting and to provide appropriate countermeasures and/or driver assistance systems, a deeper understanding of how texting distracts the driver in different traffic scenarios is required.

While NHTSA defines driver distraction as anything that diverts the driver’s attention from the primary tasks of navigating the vehicle and responding to critical events, a more

comprehensive definition is needed to better standardize and analyze exactly what is causing these collisions. Arriving at this definition is made difficult by three major factors: whether driver distraction requires an identifiable source, the question of how much control the driver has over the triggering activity, and whether distraction should include events or activities external to the vehicle as well as those inside the vehicle. Taking the aforementioned factors into consideration, a definition was presented by the Australian Road Safety Board in 2006:

- ▶ Driver distraction is the voluntary or involuntary diversion of attention from the primary driving tasks not related to impairment (from alcohol, drugs, fatigue, or a medical condition) where the diversion occurs because the driver is performing an additional task (or tasks) and temporarily focusing on an object, event, or person not related to the primary driving tasks. The diversion reduces a driver’s situational awareness, decision making, and/or performance resulting, in some instances, in a collision or near-miss or corrective action by the driver and/ or other road user.

This definition helps to separate distraction from other forms of inattention (Ranney 2008). Another definition is also offered by Regan and Lee et.al. (Lee et al. 2009) which integrates previous definitions of distraction into one broad definition:

- ▶ Driver distraction is a diversion of attention away from activities critical for safe driving toward a competing activity.

A study conducted by the Pew Internet and American Life Project estimates that the percentage of adults that said they used the text messaging feature on their phone has grown from 35% to 65% between 2006 and 2009(Madden and Lenhart 2009). Examining the process of texting while driving, one notices that every step involves one or multiple forms of distraction as shown in Table 19.4. This process inherently causes a driver to be distracted from its primary task of driving.

■ Table 19.4
Six steps of text messaging

Functions and occurrences	Type of distraction		
	Cognitive	Visual	Manual
Phone rings	X		
Check phone		X	
Retrieve message		X	X
Read message	X	X	(X)
Perceive and formulate	X		
Reply to message		X	X

() means task is possible

One difficulty for researchers attempting to quantify the effects of texting (or other distractions) while driving is the question of how you know when texting (or distraction) was a factor in causing a collision. Short of physical evidence left as a result of the collision, driver reported behavior is (unfortunately) the most reliable (and basically the only) way to know what sort of distracting activities were occurring before the collision. Accident data and police reports are not a suitable source for this purpose because people often may not admit to fault.

The National Occupant Protection Use Survey (NOPUS) is conducted annually by the National Center for Statistics and Analysis (NCSA), and provides the only nationwide probability-based observed data on driver electronic use in the United States. The survey compiled data from over 1,500 different sites (randomly selected roadway sites) involving over 55,000 observed drivers. Results of the 2008 survey included 6% of drivers seen using hand-held cell phones and 1% of drivers visually manipulating their phones while driving. Additionally, the study shows 8% of drivers age 16–24 using hand-held cell phones (compared to 6% age 25–69 and 1% age 70+) reaffirming the notion that young drivers are more likely to be using the phone and subsequently put themselves in a higher risk position. Combining the findings of this probabilistic study with those of NHTSA's 2007 Motor Vehicle Occupant Safety Survey (MVOSS) estimated that in a typical daylight moment 11% of drivers on US roadways are using a cell phone (Pickrell 2009).

The study “The Effect of Text Messaging on Driver Behavior: A Simulator Study (UK)” (Reed and Robbins 2008) shows the most important influences of text messaging while driving as well as trends in the consumer behavior. They report a poll conducted on the social networking site Facebook by the RAC Foundation discovered that of the 2002 person survey, 45% admitted to texting while driving. The foundation commissioned the Transport Research Laboratory to attempt to quantify the effect texting has on driver performance. Seventeen participants (8 male, 9 female) were recruited for the study, all between the ages of 17 and 24 and self-reported regular users of text messaging features on phones with alphanumeric keypads. The study consisted of three test drives, the first being a familiarization run and the next two being a texting drive (performing text messaging tasks – read, write, ignore) and a control drive (same route, no texting). The order in which participants completed their texting and control drives was alternated between participants to counterbalance any learning effect.

The UK study showed that text messaging causes significant increases in the mean reaction time and the number of lane departures, as well as a decrease in the mean value of the maximum speed. Besides the text messages' influence on the driving, the study also showed an influence of driving on text messaging. The time needed to write text messages increased significantly during difficult driving sections (Reed 2008).

A somewhat similar American study (Drews et al. 2009) attests the validity of the UK data for US roads and drivers. In addition, further variables like the following distance were measured. Again, increased reaction times and numbers of lane departures were found. Furthermore, text messaging was found to cause an increase in the mean following distance. While it is intuitive that a decrease in minimal following distance and increases in reaction time, standard deviation of following distance, lane crossings, lane reversals,

and gross lateral displacement are all indicators of increased crash risk, an increase in general following distance would seem to indicate the contrary. While it is possible that drivers were either consciously or unconsciously creating a “safety buffer” for themselves due to an awareness of the increased risk when texting, the buffer was hardly adequate to overcome the aforementioned risk (Drews et al. 2009).

Further validating the results of the two studies above, an Australian simulator study using novice drivers (6 months experience or less) produced similar results – significant increases were seen in the total number of lane excursions, mean standard deviation of lane position, total number of missed lane changes and mean, standard deviation, and minimum time headway. Additionally, this study utilized a device mounted in the driver’s compartment that measured the total time spent with eyes off the road, as well as the frequency and duration of these “off-road glances.” Results showed significant increases in all three metrics, further exemplifying the dangers of texting while driving (Hosking et al. 2007).

Most reported research has focused on characterization of the distraction problem and their influence on driving behavior and correlations with the measurable driving variables (e.g., lane deviations, headway, etc.). While these are very useful in assessing the problem and its effect on the driving behavior, much research is still needed to determine technology-based countermeasures for distracted driving. As mentioned earlier, these technologies must incorporate the driver condition in complex situations (e.g., in presence of multiple stimuli) and the human response capacities. Among valuable areas of investigations are those that focus on the driver cognitive and physiological capacities and performance during driving.

Electroencephalography (EEG) is the recording of voltage fluctuations resulting from ionic current flows within the neurons of the brain. EEG measures electrical activity along the scalp and has been correlated to various human conditions. Ching-Teng et. al. studied the drivers’ distraction-related EEG-dynamics in a simulator environment and concluded that dual tasks induced more event-related EEG activity in theta band (5 ~ 7.8 Hz) and beta band (12.2 ~ 17 Hz) indicating more brain resources required to accomplish dual tasks and that the increases in the theta band activity of EEG in the frontal area could be used as a distraction index for early detection of driver inattention in the future (Chin_Teng et al. 2008). Since EEG signal cannot be measured unobtrusively yet, indicators of this nature although helpful in research arena, are not yet feasible for real-life implementations.

Driver assistance technologies, along with driver education and awareness campaigns, and law enforcement, can collectively help reduce the distracted driving crashes. Many other safety problems caused by inattention can also benefit from appropriate ADAS but are still subject to research.

4 Driver Assistance Types and Levels

The goal of the driver assistance technologies is to handle as many of the arising situations as possible to improve and enhance the safety and comfort of driving. Therefore, a diver

assistance system should consider both the normal driving and a variety of hazardous situations.

Another consideration of ADAS development is to pursue areas in which the countermeasures provide the largest benefit. This requires a statistical analysis of the highway and traffic safety data. Accident data determines what situations or type of accidents are occurring and the role of the driver. However, the latter has always been very difficult to determine because accident investigation reports (police reports) cannot accurately determine the state or the action of a driver prior to a crash. Often drivers either do not remember or know what exactly happened (due to shock of trauma or the suddenness of the event) or in many instances do not admit to fault. Therefore, the cause of accident remains mostly a guessing game or a mere estimation from peripheral evidences (e.g., tire brake marks, vehicle trajectory traversed, crashes with guardrails, signposts, etc.).

The development of ADAS historically has followed the conventional causation and benefit analysis of vehicle safety systems, i.e., countermeasures have been developed to overcome most severe and most occurring situations based on the available technology, the economic feasibility, and the implementation policies to allow certain level of intervention in driving tasks.

Vehicle safety is provided through two fundamental methods of passive and active safety systems. The driver assistance system is a subset of the latter. Passive safety systems such as seatbelts, structural crush zones, and airbags have become a requirement and an integral part of a car design. Active safety systems take a preemptive role in mitigating hazards or potentially hazardous situations and aim to either avoid them all together or reduce their effect and harm in case of a collision. To this end, active safety systems also aim to overcome driver error/impairment or assist the driver in controlling the vehicle. The assistance is achieved by informing, alerting, and supporting the driver to control the vehicle. ADAS (advanced driver assistance systems) deals with the maneuvering functions of the vehicle, whereas, generally speaking the term “driver assistance” refers to all assistance types provided to the driver, including navigation guide, and the various comfort systems, etc. Therefore, active safety systems and advanced driver assistance systems serve the same goal, namely, enhancing vehicle safety and efficiency. ADAS can be considered a subset of active safety systems. (The terms “DA” and “ADAS” have been used interchangeably in the literature.) These systems augment and enhance both the driver’s perception and response.

The perception is enhanced by the use of sensors, which either perform better than humans or enhance driver’s capabilities. For example, a night vision system improves driver’s visual perception. A radar or vision-based blind spot detection system enhances driver’s capability in viewing the adjacent lanes. While in many systems that involve warning only, the response is left to the driver, other systems partially assist the driver in controlling the vehicle by augmenting the driver’s action. Therefore, ADAS could both augment an already in process response (like brake assist system) or generate a totally new response (like ACC’s automatic control of speed). For example, in a brake assist system if an impending collision is detected but the driver does not apply sufficient brake force, the system applies additional brake force to fully utilize the vehicle’s braking capacity; this is

done after the imminent hazardous situation is detected and the driver's high rate of braking is sensed. In an ACC, after the initial setting by the driver, the vehicle will automatically adjust the speed upon the emergence of speed change in the preceding vehicle. Therefore, an ADAS can be viewed as a copilot which augments both the perception and the response ability of the driver.

4.1 Classification According to Level of Intervention

Active safety systems and ADAS can be categorized according to their level of interaction with the driver or their intervention on the driving controls. For example, some active systems are employed atomically like the ESC (electronic stability control), and others provide information or warning to the drivers. At one extreme is the automatic control functions, which are vehicle dynamics-related controls that are transparent to the driver – just as an automatic transmission vehicle removes the gear shifting task from the driver of a manual stick shift car, ESC is an automatic stability control that does not require the driver intervention. At the other extreme are the pure informational systems, which simply provide useful information to the driver and enhance the situational awareness. Examples are weather advisories and information about icy road surface conditions, etc. In between these extremes are safety and comfort systems with various levels of intervention. ADAS scope mostly falls in this category. These systems could warn the driver, inform of a corrective action, or take partial (haptic) control. For example, the lane departure warning system, recently offered by several manufacturers, warns the driver of an unintended lane-departing vehicle. Many of the driver assistance systems are in the form of information and warning signals to the driver and can be disengaged optionally.

As seen in the examples above, the active safety systems can be at the informational, warning, partial (semi) control, or automatic control levels. Each level provides different types of interaction and intervention with the driver and vehicle. These are further described below. ● [Table 19.5](#) provides a general categorization of the ADAS, or more broadly defined, driver assistance systems with respect to the level of interaction or intervention with the driver.

4.1.1 Informational

As shown in ● [Table 19.2](#), the classification is basically based on the response action of the system. The informational systems, which by some researchers may not even be considered an ADAS because they do not interfere with the driving maneuver control, merely provide additional information to the drivers in a passive and nonintrusive fashion. They are not designed to alert or advise the driver of specific control actions. They merely provide useful information, which can affect the driver's decision and driving actions. Generally, they are not designed for emergency situations either. For example, information about “road icy condition” or “stopped traffic ahead” is useful for safety and comfort of driving but they are not emergency alerts or warnings.

■ Table 19.5

Active safety and driver assistance systems classification based on their level of intervention with driver

Classification of driver assistance systems	Function or task (perception: sensing, estimating, computing)	Interaction with driver or intervention in driving task (response action)
Informational	Sense environment, road, weather, retrieve real-time or archival data	Enhances situational awareness and condition monitoring: display and present the relevant information
Warning-alerting	Sense condition, evaluate situations and potential hazards, decide when and what to do, decide corrective action	Alerts the driver of potential hazards and possibly recommend corrective actions (slow down, brake, steer)
Partial (semi) control	Sense condition, evaluate situations and potential hazards, decide when and what to do, decide corrective actions	Provide both warnings/alerts and partial control functions (e.g., apply partial brake force, stiffen gas pedal to retard speeding)
Automatic (full) control	Sense condition, evaluate situations and potential hazards, decide when and what to do, decide corrective actions	Apply the vehicle control function as needed (automatically apply the brakes, ESP, etc.)
Autonomous control	Have a trip plan (from origin to destination), have navigation plan, vehicle guidance and control, sense condition, evaluate situations and potential hazards, decide when and what to do, decide corrective actions	Execute the trip plan, generate navigation, guidance, trajectory plan, and execute vehicle control; execute collision avoidance and redirection, and reroute plan and control as necessary

4.1.2 Warning

Warning systems are assistance levels that alert the driver of a specific safety condition and required action. They warn a potential imminent hazard and may actually include a recommendation signal for corrective action to be executed by the driver. The mode of warning and the persistence level are part of the design of the ADAS, which must be carefully planned while considering human factors, driver cognitive and perception capabilities, and driver behavior. The warning can be visual in the form of informative icons on the instrument panel or a head-up display, auditory such as a beeping or buzzer sound with progressively increasing frequency as needed, haptic such as seat or steering wheel vibration, and/or a combination of these. In addition to the timing of the warning, which must be in real time and immediately after sensing the hazard, another design consideration is the persistence level and intermittency of the warning signal. A nonemergency situation warrants a repeated intermittent warning like the seatbelt use

signal, whereas a crash likely situation, mandating a brake action, may require a more forceful and persuasive alert signal or intervention.

Obviously not all warnings are for an immediate danger but may alert the driver of a potential hazard. A lane departure warning merely alerts the driver that he/she is leaving the lane but does not indicate a collision. However, a sideswiping collision warning system must alert an emergency situation and a potential collision or intervene by resistance to steering.

4.1.3 Partial (Semi) Control

Up to a few years ago, most automobile manufacturers opted out of controlling the vehicle automatically and left the actual control functions to the driver. With the development of ADAS and increasing market acceptance, systems that provide control support have begun to emerge. Adaptive (or automated) cruise control (ACC) is one example of a product that partakes in the braking and acceleration control of the vehicle. The driver has the option of overriding the system by providing additional braking or disengaging entirely on the fly. The driver also has the option of setting his/her preferred thresholds for a more or less aggressive headway control.

The systems in this category provide semi-control functions that support or augment the driver actions or selected choices (like setting the ACC on cruise) automatically but do not take over the control of the car and always allow driver to overtake the control. Each of these systems addresses very specific situations, which improve the safety and comfort of driving. ACC is a comfort system for long highway driving but ensures safety as well. The more recent ACC, which can also operate in stop-and-go traffic, provides the same function in a city-driving scenario.

Similar to the informational and warning systems, systems that provide semi-control of the vehicle need to have a strong, robust, and reliable sensing capability that provides the necessary situational awareness for the function at hand, a safe decision making logic that considers the vehicle, the surrounding, and the driver input, and a robust vehicle control system and actuation mechanism which directs the (handles) the vehicle safely. These systems are truly enhancing driver's perception and response by either augmentation (in the case of brake assist) or vehicle control (in the case of ACC).

4.1.4 Automatic Control

Here automatic control refers to functions that deal with various aspects of dynamics of the vehicle and are totally transparent to the driver. They function automatically without driver intervention at all. Electronic stability control (ESC), traction control (TC), active suspension (antiroll systems), and ABS are all triggered based on certain sensed conditions of the vehicle dynamics including the vehicle interaction with road surface. They are

concerned with the servo level control of the dynamic systems (the vehicle chassis, suspension, etc.). They may provide a different ride feel or feedback to the driver but they do not interact with the driver even though they are automatically enhancing the safety of the vehicle. In this respect, many researchers or texts do not consider them a driver assistance system but rather classify them under active safety systems. Here, they are included in the broader definition of driver assistance which is a subset of active safety systems.

4.1.5 Autonomous Control

In the context of driver assistance in this chapter, full or “automatic” control is distinguished from “autonomous” control, the latter referring to systems which provide the entire function of trip planning, navigation, guidance, trajectory planning, and control, i.e., a total replacement of driver with an autopilot system for the vehicle.

Autonomous control should also provide all the decision making, situational awareness capabilities, rerouting, collision avoidance, and all required functions to successfully execute a trip from origin to destination in the presence of changes, disturbances, and uncertainties. While the autonomous driving concepts have been proven successfully in several laboratory and real-life settings, the operational safety of such vehicles in the presence of other traffic and disturbances, and possibilities of malfunctions still requires much more research and development. This subject is also treated in more depth in other sections of this handbook.

4.2 Classification According to Adaptability

ADAS can be classified according to its level of adaptability to the driver preferences, or driver behavior. There are two approaches in the design of ADAS: (1) design the systems as generic as possible and the drivers will adapt to it, and (2) design the system adaptable to individual driver's need. Both approaches have advantages and disadvantages. Of course the design of generic systems will be simpler, have broader applicability, and easier to standardize, while driver adaptable systems will require a lot more sensing capability, embedded intelligence, and knowledge of the individual driver behavior. The latter approach is still the subject of much research.

Almost all of the existing systems today are of the first type. Systems are designed to mitigate special hazardous situation or create additional standard comfort for all drivers equally. The specific problem caused by driver input is mitigated by the assistance system without any knowledge of the driver condition or behavior. A presently available lane departure warning system provides an alert when detecting merely a vehicle's lane deviation and has nothing to do with the driver's ability or behavior. The frequency and severity of warnings are solely based on preset values. Similarly, an ACC, after the initial settings by the driver, controls the vehicle's headway time or distance with the preceding

vehicle. All of these and many other ADAS examples are of generic type. They are designed based on safety first principles and available to all drivers. If public chooses to use them, they become marketable options or eventually standard on the vehicles, and drivers will adopt them.

Designing the systems adaptable to individual driver's behavior is a more daunting task. It will require much more in-depth understanding of the driver's specific behavior in response to the system or situation for which the system is designed. By definition, these systems must know a priori or learn online the driver's behavior and adjust their "assistance" function accordingly. One example will help to clarify this type of assistance. Fatigue and drowsiness is known to be a major cause of fatalities and severe injuries in vehicle crashes, particularly for road users who require long driving hours and/or night driving, etc. The significance of this problem is well established and the subject is presented in much more details in a subsequent section of this handbook. Much research has been conducted for detection of drowsiness and development of countermeasures for fatigued and drowsy drivers but a robust and reliable fault-proof system is still undeveloped. This is mainly due to the large differences of symptoms in drowsy drivers and their unpredictable responses. Any viable system must be able to accurately detect the symptoms of drowsiness in individual drivers and this has proven an extremely difficult technical challenge. Physiological measures (EEG, heart rate, etc.), facial expressions, and eye closures have been the target of research for many years. On the other hand, the driver's action and the vehicle response, e.g., steering, steering rate, steering corrections and reversals, rate of steering and combinations thereof have also been used as indicators of fatigue and drowsiness.

As seen in this example, all of these quantifiable parameters (e.g., steering reversal, steering rates, etc.) are specific to individual drivers. Thus, any detection system has to be somehow adaptable to individual drivers. Since, the response of the drivers during the fatigue and drowsiness are also highly variable, then the warning methods or control methods, perhaps need to be adaptable as well, although generic warning systems may also be developed for the majority of population.

Developments of intelligent driver assistance systems that are adaptable to individual drivers require very careful study of the cognitive as well as control aspects of driving. An intelligent copilot has to support the driver based on the ability of the driver (Michon 1993). It is envisioned that future ADAS will have a combination of both generic and adaptable support systems for the driver.

4.3 Other (Temporal) Classification of ADAS

There are two ways of looking at ADAS:

1. Support driver during normative driving
2. Support driver during emergency or presence of hazards

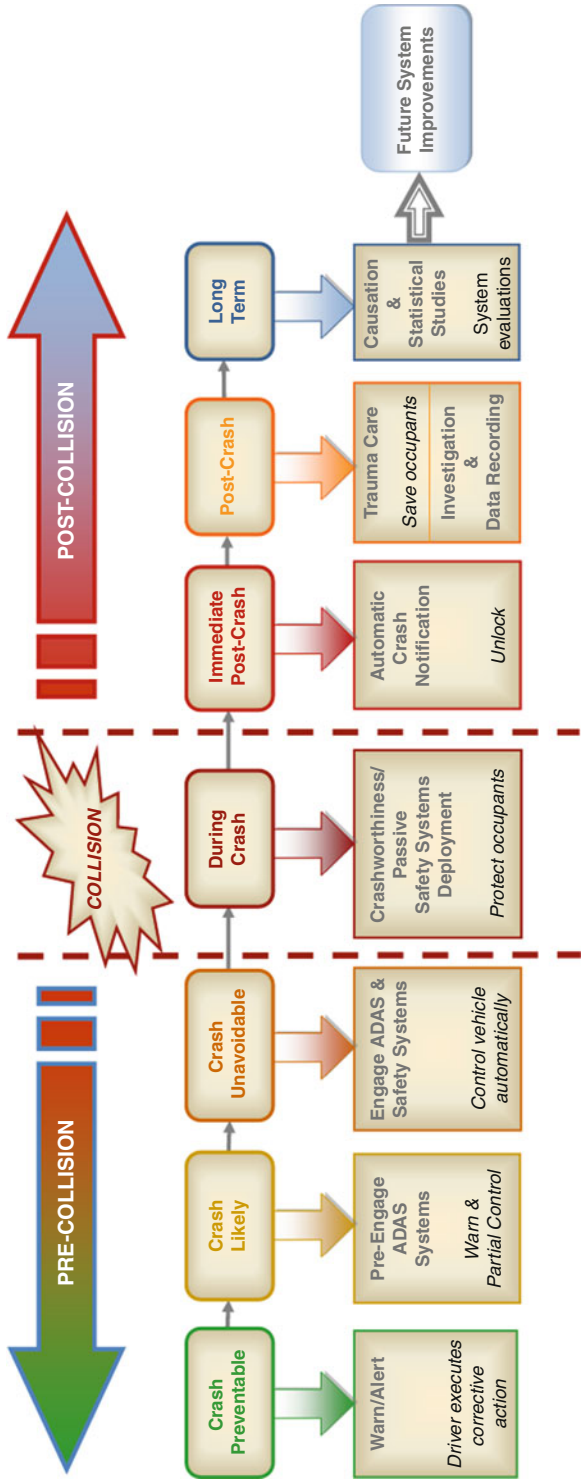
The first is primarily a comfort system with some safety implications, while the second is a critical support function of ADAS and a subset of vehicular active safety systems. The comfort systems like navigation provide better choices for the driver and, in the long term, improve both the efficiency and safety of driving. For example, navigation systems can prevent superfluous driving and hence save energy and provide alternate less congested routes resulting in a safer overall trip. The direct safety functions of ADAS aim to mitigate a hazardous situation as best as possible; these are a variety of pre-collision warning, collision avoidance, and collision mitigation systems.

In the temporal method of categorizing the active systems and ADAS, the pre- and postcrash temporal regions can divide the technologies. In this method, different safety systems or approaches are devised for different periods, i.e., moments before and after a crash or hazard occurs. A temporal division according to the conditions of *crash preventable*, *crash likely*, and *crash imminent* could correspond to launching of a different active system or ADAS for each condition. ● [Figure 19.3](#) illustrates this temporal classification of active safety (ADAS) systems.

The precrash ADAS will assist the driver to brake or steer safely around an obstacle by either warning or supporting control. The specific action taken and the ADAS deployed will be dependent on the likelihood of crash. A *crash likely* situation may be preventable by providing the warning or alert sufficiently early for driver to take the necessary action. A *crash likely* ADAS may create both warning and support partial control. A *crash imminent* situation may warrant an ADAS to take full control as the alternative would be even more severe crash pulse and injuries. In this case, a full braking or complete swerving (evasive maneuver if situational awareness warrants) will be the most appropriate. In temporal classification, the level of intervention and the system action are similar to the earlier defined classification.

Another group of ADAS are deployed for postcrash injury mitigation. These systems aim to provide the most appropriate emergency care for the situation at hand and typically deal with high severity crashes. Crash occurrence and perhaps crash severity is automatically detected and transmitted to dispatch centers, which then inform the corresponding authorities simultaneously, e.g., police, fire department, ambulance, etc. The present systems like the GM's "On-Star" system ([Web site Onstar System](#)) and other similar roadside services, provide postcrash alerts to police, ambulance, etc., upon automatic remote detection of a crash. The postcrash medical services can have life-changing effects every minute after a severe injury occurs.

The present systems are limited to reporting the accident and associating severity by detecting, e.g., if the airbag was deployed, or the maximum velocity at impact, the number of impacts, and rollover status. This measures the impact, and helps first-responders gauge the severity. An advisor shares this potentially life-saving information with responders before they arrive. However, research is underway for development of urgency algorithms, which provide additional parameters correlated with the severity of crash in automatic crash notification (ACN) systems (Augenstein et al. 2001, 2007). The crash type and angle, crash force or acceleration pulses from different sensors, and deployment of safety systems (belt



■ Fig. 19.3 Illustration of precrash, crash, and postcrash temporal relationship with vehicle occupant safety systems

pre-tensioners, airbags, etc.) all play a role in the resulting occupant injuries. Establishing simple relationships and correlations of crash conditions with injury mechanism is a complex engineering and medical problem and subject of research. Knowing more details of possible injuries (head, chest, lower limbs, lacerations, etc.) and injury levels is critical in life-saving ability of the postcrash trauma care for the injured occupants.

5 Integrated Safety

A discussion of ADAS is not complete without a brief mention of *integrated safety*. Integrated safety considers all available safety tools to mitigate hazards and reduce harm to the driver and vulnerable road users (bicyclists and pedestrian).

Various passive and active safety systems are integrated automatically to provide greater benefits as part of a vehicle safety strategy. For example, restraining belts are equipped with pre-tensioners, which could be triggered early on during a crash imminent condition, sensed by the vehicle collision avoidance radars. The application of pre-tensioners will better secure the occupant to the seat, prior to the crash and hence minimize the risk of injury. This is but one simple example of a more general approach to safety integration strategy.

In a holistic approach to integrated safety, sensors provide added condition monitoring and situational awareness. Safety algorithms use the sensory data and apply various deterministic and stochastic methods for assessment, evaluation, estimation, perception, and decision making. These algorithms provide the optimum and safest corrective actions, i.e., decisions that an expert driver would make at a critical situation. These decisions, acting as a smart copilot, are then translated to specific vehicle controls (braking, steering). The actual function of interaction/intervention/actuation is part of the design of the safety system. It may be executed at a different level as listed in the above table, i.e., span from information and alerts to full controls.

Inter-vehicular and vehicle to infrastructure communications are bringing the next wave of vehicle safety advancements. Some OEMs have demonstrated the vehicular communications to warn drivers of intersection collisions and to minimize the vehicle blind spots, among other applications. Dedicated Short Range Communications (DSRC) and other communications protocols are being evaluated for various safety applications. US DOT has been conducting field operational tests and collaborative research under the ITS program (cooperative driving; previously *IntelliDrive*, Intelligent Vehicle Initiative, etc.) to advance the future standards. Many other integrated safety systems will be forthcoming: lane departure control, run-off-road warning and control, assisted braking, blind spot warning, intersection collision warning, intelligent speed adaptation systems, drowsy/fatigue driver detection and warning, impaired driver assistance, among many others, are either available today or are in the works.

Undoubtedly, the advanced integrated safety systems have proven very effective in reducing the harm in prototype and laboratory settings, and will benefit the users. The major issue concerning these active technologies is the unavailability of suitable uniform

standards of performance. NCAP type testing (NHTSA's New Car Assessment Program), which provides crash standard ratings for vehicles, are now needed for assessment of these new active technologies.

Various standard crash imminent test scenarios are presented in (Ference et al. 2007) for the purpose of objectively verifying the performance of integrated vehicle-based safety systems. The scenarios were obtained from the GES crash database. The authors discuss the scenarios application for the verification of integrated vehicle-based safety systems with a focus on rear-end, lane change, and run off-road crashes.

Devising suitable standards for active safety systems is a complex problem due to variations of the packaged systems, nonuniformity between manufacturers, and the complex nature of the systems which interact with the driver, hence adding the driver's adaptability issues to the problem. Much research and evaluation is still needed to arrive at innovative and effective standards for these technologies.

6 Human–Vehicle Interface

As described above the ADAS can be classified according to the level of intervention with the driving task. Any intervention, whether a simple icon or light in the instrument panel or a complex steering assist, could potentially have a negative impact on driver's attention on the road. The challenge on the design of ADAS is to minimize this impact. The principles of human–machine interface apply to many aspects of driving and must be considered in human factor designs of in-vehicle systems including ADAS. Research has shown that the negative impacts of driver assistance systems are particularly distractive for elderly drivers (Hall and Dudek 1994).

HMI for ADAS must accommodate three modes of exchange between the driver and the machine (vehicle):

1. Provide feedback to driver
2. Receive driver's input
3. Execute automatic (and autonomous) action (actuations)

Each mode is explained briefly.

6.1 Feedback to Driver

Providing feedback to driver is through receptive human senses. They can be

- Visual
- Auditory
- Haptic
- Olfactory

Each of these has a different perception time. Many other factors listed earlier and the complexity and the number of tasks at hand also affect the perception time. Drivers' visual

sensing can be triggered by all other senses as well as visual stimuli. Response elicited from stimuli that can be received by the visual (eye), auditory (ears), and tactile senses can result in additional visual searches by the driver for the identification stage of perception (Olson et al. 2010, pp. 378). For example, the appearance of an informative icon on the instrument panel, or an unexpected audio warning, could cause a visual search by the driver who is not used to the signal to identify the source of fault or understand the instructions. This identification could be a source of difficulty and lengthening of the PRT.

All input stimuli, including visual, auditory, haptic, and cognitive contribute to a driver's workload. Drivers have some visual space capacity that allows them to look away from the road ahead for certain periods. This will create a visual diversion but not necessarily distractions. Visual distraction occurs when this capacity is exceeded by a stimulus, i.e., drivers focusing too long on a specific task, or fixation on an object.

Auditory capacity of drivers may be even higher, as drivers can listen to radio or messages without creating driving hazards, although it is possible to imagine situations in which audio messages could create a mental load or cognitive overload. Cognitive overload may result when driver thinks about things other than driving. This is the most difficult capacity to measure or quantify. Manual overload occurs when a driver undertakes other secondary physical manipulation tasks during driving such as drinking coffee, adjusting radio stations or other vehicle controls. The existing driver assistance systems try to minimize this workload by creating easier to manipulate controls on steering wheel for volume adjustment, cruise control settings, etc.

Haptic feedbacks are part of the standard vehicle design to allow control of the vehicle. Steering feel or response is a characteristic of vehicle design. The responsiveness of gas and brake pedals is also of haptic nature. Vehicle ride comfort depends on vibration feedback to the driver through the structure and seat. Luxury sedans provide a plush feeling of a smooth ride whereas more sporty vehicles usually provide stiffer structural and vibratory responses. Haptic systems can provide a warning feedback to the driver in place of, or in combination with audio and visual signals. For example, seat or steering wheel vibrations can be used as virtual rumble strips to warn drivers of unintended lane departures. Highway rumble strips have been very effective safety features, and a similarly emulated effect is anticipated to provide comparable safety results for center lanes departure warning. The effects of rumble strip vibrations on driver and seat has been studied in depth by Laoufi and Eskandarian (Laoufi 2005).

Olfactory feedbacks are not among currently used feedback mechanisms. Ho and Spence provide empirical demonstration that olfactory stimulation can facilitate tactile performance (Ho and Spence 2005). Research has been done on triggering this sense for alertness, e.g., to keep drowsy drivers alert or create a more positive feedback to increase wakefulness but no substantial definite success has been reported (Mallis et al. 2000).

Another consideration in design of in-vehicle safety systems is the distribution of visual (and audio) and manual tasks expected of the driver. Drivers could demonstrate time-sharing strategy for visual task with having the capability of peripheral vision and a broader field of view, but manual tasks occur sequentially.

Integrated warning modes can be useful if not overwhelming the driver. Haptics can augment the audio/visual feedbacks without adding a cognitive overload in many circumstances. Integrated warning needs to consider the physical, perceptual, and cognitive needs of drivers. There are no hard rules for design of effective warning systems and there are no design principles which will ensure that a warning will be effective (Laux and Mayer 1993). Laux and Mayer offer some general guidance on questions to ask when designing a warning system (Laux and Mayer 1993, p. 415), but emphasize that the effectiveness of the warnings can only be evaluated partially in field testing and through the users interaction with the vehicle.

6.1.1 Age Consideration

Age has proven to play a role in PRT and in the response reaction of the older drivers. ADAS must ensure safe operation by older drivers. Enumerating all the differences and issues concerning elderly drivers is obviously beyond the scope of this chapter. Transportation Research Circular 419 (Hall and Dudek 1994) highlights many issues concerning driving performance of older drivers and IVHS (Intelligent Vehicle Highway Systems). According to this report although both younger and older drivers are functionally equivalent in the performance of predetermined single control movement responses, older drivers show significant decrement in performance of two and three movements. This means older drivers' sequential psychomotor response times are higher than young drivers and that they would be at higher risk when a sequence of control movements are required (e.g., during an emergency evasive maneuver immediately following another movement.) (Hall and Dudek 1994, pp. 17–18)

6.1.2 Mental Workload

Driver mental workload is affected by all external and internal stimuli in addition to the demand required by the driving function. The following factors affect cognitive and mental workload:

- Traffic
- Bad weather
- Poor visibility and lighting
- Poor surface conditions
- External noise
- In-vehicle noise
- Poorly or partially functioning systems (heater/AC, windshield wiper)
- Poorly responding vehicle dynamics (acceleration, braking, and steering responses)
- Lost directions
- Sources of anxiety (urgency of trip, children crying, etc.)
- Driver's health and mood

The purpose of ADAS in addition to assisting the driving function should also be to minimize the above negative mental workload effects. Many systems are already in place, which do provide this service. For example, timely weather and road condition advisory system may provide the necessary advice and warnings during trip planning or en route, hence, better prepare the driver by building the expectations. Proper diagnostic systems could minimize malfunctioning by early service warnings. The automatic active safety systems (e.g., stability control, traction control, etc.) help mitigate driving on poor road surface and weather conditions. However, accommodating all workloads or a combinatory effect of multiple factors still requires much development effort.

Although the intention is to minimize the effects of in-vehicle systems overload on the driver, many of the input sensory stimuli and some of the sources of additional mental workload inevitably do occur during normal driving and may be increased by a new ADAS. The ADAS produced workload should not exceed the driver's visual, auditory, and haptic and cognitive capacities. Complex designs with poorly designed feedback mechanism will diminish the safety value of ADAS and the acceptability of the system.

6.1.3 Vigilance

Vigilance refers to how much attention the driver is paying to the primary task of driving. In normal (and vigilant) driving, this manifests itself primarily in the driver's looking at the road ahead continuously with the appropriate attention. Deviations from this posture or condition determine a degree of lack of vigilance. The degree of vigilance of a user can be related to the open or closed state of his/her eyes and mouth and to the frequency of his/her blinking and yawning. Many factors such as fatigue, drowsiness, and external disturbances could cause degradation of vigilance. For example, fatigue and drowsiness cause a higher percentage of eye closure, measure by a variable PERCLOS (Wierwille 1999). PERCLOS refers to the percentage of time that the driver's eyes are between 80% and 100% closed during a defined time interval. This variable is an indication of unsafe driving when the pupils are covered for sufficiently extended percentage of time and do not perceive visual stimuli.

Other effects include fixation of eyes on objects, frequent blinking, head nodding, or unnatural facial features indicating fatigue. Measuring vigilance or having standards for what constitutes appropriate level of vigilance is difficult and varies among drivers. But fortunately, there is some progress in this area. Eye tracking and computer vision technology has made great strides in being able to measure some of the facial and eye features accurately and correlate them with vigilance.

Sigari (2009) developed an algorithm for driver hypovigilance detection based on eye-region spatiotemporal processing without explicit eye detection stage. Symptoms of fatigue and distraction, including percentage of eye closure (PERCLOS) and eyelid distance changes were used for fatigue detection; and eye closure rate was used for distraction detection. Bergasa et al. (2006) developed a real-time system with some success for monitoring driver vigilance which took a combination of six parameters: Percentage of

eye closure (PERCLOS), eye closure duration, blink frequency, nodding frequency, face position, and fixed gaze. They use a fuzzy classifier to infer the level of inattentiveness of the driver. They report the combination of parameters yielded more accurate measure of vigilance than using any one parameter alone.

Sommer et al. (2009) proposed an independent reference of driver's hypovigilance is needed for the assessment of fatigue monitoring technologies. They process EEG and EOG biosignals to apply a feature fusion concept and to utilize support vector machines (SVMs) to determine two classes of "slight" and "strong" hypovigilance. The test results of 16 subjects in a driving simulator experiments were compared with PERCLOS (percentage of eye closure), and oculomotoric variable utilized in several fatigue monitoring systems. They conclude that EEG and EOG biosignals contain substantial higher amount of hypovigilance information than the PERCLOS biosignal.

6.2 Receive Driver's Input

The ADAS interface receives drivers' input and either executes as instructed or augments it as necessary. Presently the inputs for the driving tasks are the usual in-cabin control functions, e.g., brake or steer. The augmentation could be the brake assist, e.g., vehicle receives initial brake and if sensed inadequate, then it provides additional brake force.

For secondary tasks and comfort systems, ADAS may use manual or voice inputs, or both. Many hands-free phones and navigation systems now are voice activated. They include a user interface to be trained and learn the voice of the driver and minimize command errors.

6.3 Actuations: Execute Automatic and Autonomous Action

This part of man-machine interface handles the actuation task. This is the task that needs to be accomplished by the vehicle after the driver input is received or automatically without driver's intervention. The subject of autonomous driving is covered at length in another section of the handbook and not elaborated here. It suffices to mention that two main levels of decision and control are required. The higher level autonomy provides human-like decision making for situational awareness, identification, and planning, which sends commands to lower level servo controls to actuate the vehicles acceleration, brakes, and steering. ADAS for autonomous (driverless) driving must provide these capabilities.

If in an ADAS, an augmentation of the driver input is needed then the automatic system must work over the existing manual system (or in conjunction with them) to override or extend the manual system. A steering assist has its regular steering mechanism and utilizes an additional motor to generate an overriding torque for corrections. If always a purely automatic function is needed for an assistance task, then the manual could be eliminated; this is not a scenario in regular driving environment but is suitable for robotic vehicles.

In Germany, research has been conducted to maneuver vehicle by cognitive inputs by the driver's thoughts of turning right or left (or thinking about the right side or left side). Although very far reaching, robotic vehicle control by thought has been demonstrated in laboratory settings ([Web AutoNOMOS Labs, Germany](#)). This is an ultimate form of man-machine interface, which requires much pondering.

7 Testing and Evaluation

As there are significant variations in drivers' capabilities, a driver assistance system must be thoroughly tested, at least for an average population with normative behavior. What constitutes normative driving behavior is the subject of much in-depth investigation by itself but here the normative refers to a standard or simple majority population in handling specific situations. Despite this definition, even the normative behavior may differ regionally. The following are among several factors that affect the normative behavior.

- Driver education
- Driver background
- Driver training
- Driver acceptance
- Drivers' perceptions of the safety benefits

Realistic testing of ADAS must distinguish between the countermeasure's (the ADAS) perceived versus true effectiveness. While the perceived effectiveness may be useful for product marketing and create a false image of safety or comfort for the driver, it will not truly enhance vehicular safety.

ADAS reliability is at the heart of its acceptance and eventual effectiveness to enhance safety. Systems with too many false alarms are rendered useless and not used by the drivers. Systems with mostly correct alarms but some missed warnings may have a better chance of acceptance. The drivers may not rely on them 100% but utilize them as an auxiliary safety measure, i.e., when they do issue an alarm, the driver will respond. This may indeed add to driver vigilance, whereas the same system with 100% success, while removing the burden from the driver, may also reduce the vigilance by creating a trusted safety perception. Therefore, some argue that in cases of not total autonomy in which the driver is expected to keep continuous control and maintain vigilance, a high percentage of success of detecting hazards may be preferred over 100% detection rate. The non-perfect system will have a positive behavioral effect which would increase vigilance (and less reliance on the system) and enhance safety. Although sensible, this claim needs to be investigated on a case-by-case basis.

After conceptualization, design, and prototyping, ADAS can be evaluated in four steps. First, the system can be examined in a driving simulator. Care must be taken in designing the simulator experience as close to real life as possible, ensuring the safety systems realistic emulation, along with all other traffic and environmental conditions.

Second, track testing can provide some level of functionality evaluation and verification but lacks the existence of other external stimulus from real traffic. Third, field operational test and naturalistic driving data collection are a natural addition to the evaluation process, but at a major cost. Instrumented vehicles are given to a population of volunteer subjects to drive over an extended period and data are gathered on several variables of interest. The accumulated data are compiled systematically and analyzed extensively through a variety of methods including data mining, feature extractions, and other analysis techniques of large datasets. Statistical analysis of desired variables, determine the reliability, effectiveness level, and safety implications of the ADAS.

Finally, the true evaluation of the systems takes place after real-life deployment of the system in the market and monitoring their effects over an extended period. Just like the effect of any countermeasures for safety (seatbelts, airbags, ESC), statistical analyses of accident data, causation studies, and consumer surveys, etc., determine the true value of an ADAS. This is a long-term evaluation process which requires patience and resources.

8 Samples of Driver Assistance Systems (Case Studies) and Future Research Needs

This section reiterates some of the systems described throughout this chapter and covers some additional currently available ADAS and their functions in supporting the driver. This is not intended to be a comprehensive list but rather a sample of available systems which highlight the significance of the issues discussed in this chapter. Other sections of this handbook cover several safety systems in intelligent vehicles today. This chapter is concluded by a brief list of important research in driver assistance.

8.1 Examples of Existing ADAS

The following are already in the market or implemented in prototypes:

1. Adaptive cruise control (ACC)
2. **Adaptive headlights**
3. Anti lock breaking system (ABS)
4. Autonomous driving
5. **Brake assist system**
6. **Blind spot detection**
7. Cooperative driving
8. Driver drowsiness monitoring and warning(DDMW)
9. Electronic stability control (ESC)
10. Gear shift indicator
11. **Lane departure warning(LDW)/lane departure control**
12. Lane change assistant(LCA)
13. Night vision(NV)

14. Obstacle and collision warning(OCW)
15. Steering control-evasive maneuver
16. Pedestrian/vulnerable road user protection
17. Tire pressure monitoring system (TPMS)

From the above list, the automatic active safety systems that do not interface with the driver are not discussed here (e.g., ABS, ESC, etc.). Also safety systems which are covered in other sections of this handbook are not included here. A few from the above list (shown in bold) are described in more details for illustrative purposes even though they may be covered elsewhere in this handbook.

8.1.1 Brake Assist System Supports Braking by Intervention

Automatic braking or brake assist system (also known as advanced emergency braking systems) is a new technology that ensures that the maximum pressure is applied by the brakes to stop a vehicle in an emergency situation. Some manufacturers also refer to the same system as brake assist systems (BAS; Mercedes-Benz) and emergency brake assist (EBA). While the majority of people press the brake pedal rapidly in an emergency situation, they often do not do so with sufficient force (Breuer et al. 2008). BAS enables utilization of vehicle's full braking potential.

This technology is used for longitudinal control of vehicle as an active safety system to mitigate rear-end collisions, collisions with objects, and collisions with pedestrians and cyclists, among others, where emergency braking might be needed.

Rear-end collisions emerge among the most common type of crashes, accounting for 31.5% of all crashes (Traffic Safety Facts, NHTSA 2009). Those accidents frequently represent a breakdown in driver longitudinal control and hazard detection (Lee 2006).

Brake assist system supports the drivers who under deploy the brakes in an emergency situation. Based on findings regarding the inability of most drivers to use the full potential of the vehicle brakes the system applies full brakes when a quick brake reaction by the driver is detected. Systems may use a different method as an indicator of emergency situations. It may detect an emergency braking condition based on the *brake pedal force*, *brake pedal speed* applied by the driver or by *multiple criteria*, one of which must be the rate at which the brake pedal is applied.

8.1.2 Adaptive Headlights Enhances Visibility and Safety at Night Driving

Enhanced night visibility has a direct correlation with safety, especially in handling emergency situations. Visibility is reduced in negotiating turns as the light points toward vehicle centerline ahead rather than to the direction the vehicle is heading. Adaptive headlights use swiveling headlights that always point in the direction the vehicle is steering. Consequently, the road ahead is better illuminated and obstacle visibility is

improved. A control unit uses sensors that measure speed, steering angle, and yaw angle (rate), and turns the headlights by small electric motors such that the light beam falls on the road ahead. The adaptive headlight is deactivated in the event of the vehicle oversteering/understeering or yawing (which can be detected by the steering angle sensor and the yaw rate sensor.)

8.1.3 Blind Spot Detection Augments Driver's Field of View and Provides Visual Feedback

Blind spots are areas around the vehicle for which driver does not have a clear view (directly or in the mirrors), and hence are sources of hazard in lane changing, merging, etc., maneuvers. Blind spots are areas in adjacent lanes of traffic that are blocked by various structures in the automobile. The physical constraints in eye movement and head and body rotation make certain areas invisible to the driver. The direct blind spots are areas covered by the A-pillar (between the front door and windshield), B-pillar (behind the front door), and the C-pillar (ahead of the rear windshield.) The indirect blind spots are the regions between the driver's peripheral vision on the sides and the area that is covered by the rearview mirror. The blind spot for trucks and buses are much larger than those of passenger cars. Areas directly to the right of the cab extending past the trailer, directly behind the trailer, to the immediate left of the cab, and directly in front of the cab are blind spots for truck drivers.

Proper adjustment of the center and rearview mirrors can eliminate the blind spots ([Inficon Website](#)) in most passenger cars but unfortunately most drivers are not familiar with the method and often their personal mirror setting creates blind spots.


In blind spot detection, radars mounted on rear bumper or vehicle sides scan and monitor the blind spot region. The system uses detection algorithms, tracking, object-to-lane mapping, and provide feedback to the driver (illuminating warning LEDs), to inform of the presence or absence of objects in the blind spot region. For example, by viewing the side-view mirror and observing the LED colors, the driver can decide if it is safe to change lanes. These systems have been around since 2005, but were included exclusively in higher end vehicles. They are appearing more in recent vehicles as options.

8.1.4 Lane Departure Warning Improves Situational Awareness, Warns/Alerts driver/Lane Departure Control Improves Situational Awareness, Augments Steering Control

Lane departure warning systems alert the driver of unintended lane departures. Lane departure control actually assists to bring the vehicle back into the lane. These are safety systems that support run-off-lane crashes. The National Highway Traffic Safety Administration (NHTSA) has long recognized that single-vehicle road departure (SVRD) crashes lead to more fatalities than any other crash type (Wang and

Knipling 1994). Lane departure warning (LDW) was a key technology identified at the start of the Intelligent Vehicle Highway System (IVHS) program that could potentially reduce the number of fatalities and injuries associated with SVRD (Mironer and Hendricks 1994).

A variety of systems are offered by different manufacturers which are based on speed and the monitoring and detection of lane markings. Video sensors (mounted behind the windshield, typically integrated beside the rear mirror), laser sensors mounted in the vehicle front, or infrared sensors (mounted either behind the windshield or under the vehicle) scan the road to detect an unintended lane change (e.g., a lane change without turn signals), then the system warns the driver by haptic vibration or beeping audio signals. In the case of lane departure control, the system assists to steer (without taking over control) to bring the vehicle back into the lane subtly, but the driver is still in control of steering.

A few examples of existing lane departure warning and/or control systems are shown in  Table 19.6.

Many other safety advances and collision avoidance technologies are discussed in the literature (Vahidi and Eskandarian 2003) and with the increasing power of electronic,

 Table 19.6

Sample of existing lane departure warning systems in the market (not a comprehensive list)

Year	ADAS	Company	Vehicle
2001	Lane keeping support system	Nissan	Cima; Japan (Ivsource Website)
2002	Lane monitoring system	Toyota	Cardina, Alphard Japan (Japanvehicles Website)
2003	Lane keep assist system	Honda	Inspire (Honda Website) (worldcarfans Website)
2005	Lane departure warning system	Citroën	2005 C4 and C5, and C6 (psa-peugeot-citroen Website)
2006	Multimode lane keeping assist	Lexus	LS 460 (wikipedia Website)
2007	Lane assist	Audi	2007 Q7 (audiworld Website).
2007		BMW	BMW 5 and 6 series (wikipedia Website)
2008	Lane departure warning	General Motors	2008 Buick Lucerne (wikipedia Buik Website)
2008	Lane departure warning system same as citroën	Peugeot	Peugeot 308 (wikipedia Website)
2009	Lane keeping assist	Mercedes-Benz	2009 E and S-class (Daimler Website).
2010	Lane departure warning system	Kia	2011 Cadenza premium (Autonews Website)
2011	Lane keeping assist	Ford	Ford select models 2011 (Ford Website)

radars, and digital cameras, more systems will be emerging in the market. The following is a final example to demonstrate this point.

8.1.5 Driver Assistance Pack

Ford offers the driver assistance pack as an extra option on the Titanium and Titanium X models. It is designed as a safety system to help the driver with more information and safety aids. It has different components: **Traffic sign recognition**: A front-mounted camera constantly scans the road ahead and recognizes both speed and overtaking signs, and displays both on the dash. The notifications appear large on the dashboard with a changing fading color so when the speed limit changes it is immediately obvious. **Low Speed Safety System** (in UK) and **Active City Stop** (in the United States): designed to automatically prevent low-speed collisions that can occur with urban driving. A forward-facing infrared laser beam mounted next to the camera in front of the rearview mirror scans the distance to an approaching reflective object, and upon sensing a collision, it will apply the brakes. It will completely stop the car with no driver intervention at speeds less than 10MPH, it will slow the car down at speeds between 10MPH and 20MPH, and it does not operate at speeds above 20MPH ([Expert Review Website](#)).

8.2 Future Research

Future research in driver assistance can involve numerous areas including but not limited to driver research, vehicle research, driver–vehicle interface research, and all in conjunction with the external environment, namely driver–vehicle research in the context of traffic. As vehicle research by itself is technologically driven, the most difficult challenge is the driver role in interfaces with vehicle which deals with driver behavior, controls, and physiological and psychological capacities, all of which are influenced by many other factors such as education, experience, age, etc.

McAdam (2003) examined the role of the human driver from a control modeling perspective within the traditional driver–vehicle system and proposes several important considerations for integrated human–vehicle control models, including several required research areas some of which are listed here due to their direct relevance to ADAS:

1. Modeling driver skill in analyses and computer models
2. Understanding interactions of smart vehicles with human drivers in control of the vehicle
3. Characterization of novice and less experienced driver control behavior
4. Methods for modeling human conceptualizations of how drivers “internalize” their view of the external world/controlled vehicle
5. “Development of truly integrated lateral and longitudinal control behaviors of the human driver, in a more unified and natural manner, rather than as a collection of

scattered models that represent only specific instances of these types of driver control behavior”

6. Validation and suitable measurements of human–vehicle interactions under a wide variety of operating conditions

The following can also be added to the above list as more far reaching research needs of ADAS in the future:

7. A more in-depth understanding of cognitive and control capacities of individual drivers
8. Development of learning methods for online characterization of the above item 7
9. Models that can represent drivers cognitive and physiological control limits and augment it in parallel or complementary actions
10. Development of testing standards and procedures for the evaluations of driver assistance copilot systems
11. Development of comprehensive autonomous decision analysis system to overtake control of vehicle when enhanced safety is ensured, which are applicable to most commonly encountered situations, and are adaptive to new ones
12. Modeling and rating systems, which monitor human driving behavior and provides off-line feedback for corrective actions (automatic driving tutor for novice drivers)

Several research tasks which are subsets of the above can be envisioned. It is hoped that with an appropriate vision and policies in place, research and technological advancement will enable ADAS to reduce traffic fatalities to near zero and minimize the risk of driving, while enhancing the comfort of driving and quality of life for the motoring public.

9 Concluding Remarks

The goal of driver assistance (DA) systems is to assist drivers in the primary and secondary tasks of driving to enhance safety, comfort, and efficiency of vehicles. The primary tasks are those that involve the actual control of the vehicle and the secondary tasks are the other in-vehicle operations. The term “ADAS” (advanced driver assistance systems) refers mostly to those functions that improve the handling and maneuvering of the vehicle. Several driver assistance systems (and ADAS) have been developed by the automobile manufacturers and are available in market now.

The design of driver assistance systems must have a human-centric approach, which considers the perception–response capabilities and other conditions of the human during driving. It also has to take into account the various man–machine interface principles. Driver assistance systems can be categorized according to their level of intervention with the driver and the vehicle. Informational and warning type assistance leave the vehicle control tasks to the driver, whereas partial control, automatic, and autonomous assistance take an active role in controlling the vehicle braking and steering. The higher levels of intervention require more robust, reliable, and accurate sensing and actuation, as well as higher level of intelligence and decision making, or otherwise, they will not be practical. Fortunately, with the advances in sensors, computing, and automatic control, several

vehicle dynamics functions like traction, stability, and engine control, and automatic braking are successfully deployed. However, autonomous functions that require human-like situational awareness and take over the vehicle control, at least partially, in response to surroundings (e.g., evasive maneuvers, etc.) are much more difficult to implement and still subject of further research, although much progress has been made with prototype vehicles in laboratory settings and demonstrated in closely monitored field tests (various autonomous driving demonstrations).

While generic ADAS (like ACC, blind spot detection, LDW, etc.) that are applicable to a wide variety of driver behaviors are easier to design and implement, systems that need to be adaptable to individual driver's traits are much more difficult to develop due to the variability in human perception–response and driving behavior. Human conditions of fatigue, drowsiness, distractions, and impairment further complicate the design of safety systems. Drastic loss of control is associated with some of these conditions, in most cases resulting in fatal or severe injury crashes. Driver's responses and reactions under these conditions are not well characterized or understood. Postcrash driver assistance with automatic crash notifications and deliverance of the most suitable rescue and trauma care for effective treatment of injured occupants (and vulnerable road users) should also be in the forefront of priorities for assistance systems.

Among the most important challenges of ADAS development are those involving more in-depth and comprehensive understanding of the human cognitive and driving behavior for systems, which need to cooperate seamlessly (as an automatic copilot) with the driver in the loop. Other challenges include the autonomous systems, which require human-like decision making in interaction and response to various traffic conditions and unexpected situations that may arise both external and internal to the vehicle. Research in these areas is needed to advance the next generation of driver assistance systems. As more ADASs are emerging in the market, testing and evaluation methods and international standardizations are also required to ensure safety, some level of uniformity, and public acceptance of the systems.

References

-
- Ascone D, Lindsey T, Cherian Varghese C (2009) An examination of driver distraction as recorded in NHTSA databases. NHTSA research note, DOT HS 811 216, Sept 2009, pp 1
- Augenstein J, Digges K, Ogata S, Perdeck E, Stratton J (2001) Development and validation of the urgency algorithm to predict compelling injuries. In: The 17th international technical conference on the enhanced safety of vehicles (ESV). Amsterdam, 4–7 June 2001
- Augenstein J, Digges K, Perdeck E, Stratton J, Bahouth G, Peter Baur P, Messner G (2007) Application of acn data to improve vehicle safety and occupant care. Paper number 07-0512. In: The 20th international technical conference on the enhanced safety of vehicles conference (ESV) in Lyon, France, 18–21 June 2007
- Behringer R, Sundareswaran S, Gregory B, Elsley R, Addison B, Guthmiller W, Daily R, Bevely D, (2004) The DARPA grand challenge – development of an autonomous vehicle. In: IEEE intelligent vehicles symposium, 14–17 June 2004
- Bergasa LM, Nuevo J, Sotelo MA, Barea R, Lopez ME (2006) Real-time system for monitoring driver vigilance. IEEE Trans Intell Transport Syst 7(1):63–77

- Breuer JJ, Faulheber A, Frank P, Gleissner S (2008) Real world safety benefits of brake assistance systems. *ESV* (2008) Paper number 07-0103
- Broggi A, Bombini L, Cattani S, Cerri P, Fedriga RI (2010) Sensing requirements for a 13,000 km intercontinental autonomous drive. In: 2010 IEEE intelligent vehicles symposium. University of California, San Diego, 21–24 June 2010
- Broggi A (2011) Paving the road to autonomous driving: a new challenge in vehicular robotics. Plenary talk. In: IEEE international conference on robotics and automation. Shanghai, 9–13 May 2011
- Campbell BN, Smith JD, Najm WG (2003) Examination of crash contributing factors using national crash databases. National highway traffic safety admin., technical report, DOT-VNTSC-NHTSA-02-07, Oct 2003
- Card SK, Morgan TP, Newell A (1983) The psychology of human-computer interaction. Lawrence Erlbaum, Hillsdale, pp 23–97
- Chin_Teng L, Hong-Zhang L, Tzai-Wen C, Chih-Feng C, Yu-Chieh C, Sheng-Fu L, Li-Wei K (2008) Distraction-related EEG dynamics in virtual reality driving simulation. In: IEEE international symposium on circuits and systems, ISCAS 2008, pp 1088–1091
- Dingus TA, Klauer SG, Neale VL, Petersen A, Lee SE, Sudweeks J, Perez MA, Hankey J, Ramsey D, Gupta S, Bucher C, Doerzaph ZR, Jermeland J, Knipling RR, (2006) The 100-car naturalistic driving study phase II – results of the 100-car field experiment, DOT HS 810 593, Contract or grant no.DTNH22-00-C-07007. April 2006
- Drews FA, Pasupathi M, Strayer David L, (2004) Passenger and cell-phone conversations in simulated driving. In: Proceedings of the human factors and ergonomics society 48th annual meeting. New Orleans, pp 2210–2212
- Drews FA, Yazdani H, Godfrey C, Cooper JM, Strayer DL (2009) Text messaging during simulated driving. *Hum Factor: J Hum Factor Ergon Soc* 51:762–770
- Ference J, Szabo S, Najm WG (2007) Objective test scenarios for integrated vehicle-based safety systems. Washington, DC. http://www.umtri.umich.edu/content/obj_test_scen_ivbss.pdf
- Green M (2000) How long does it take to stop? Methodological analysis of driver perception-brake times. *Transp Hum Factor* 2:195–216
- Green M (2009a) Driver reaction time. <http://www.visualexpert.com/Resources/reactiontime.html>
- Green M (2009b) Perception-reaction time: is Olson (& Sivak) all you need to know? *Collision. The International Compendium for Crash Research* 4(2): 88–93
- Godthelp H, Farber B, Groeger J, Labiale G (1993) Driving tasks and environment. In: Michon JA (ed) *Generic intelligent driver support*. Taylor and Francis, London pp 21
- Hall J, Dudek CL (1994) Committee on user characteristics. In: *Driver performance data book update: older drivers and IVHS*. Transportation research circular 419. National Research Council, Washington, DC
- Hick WE (1952) On the rate of gain of information. *Q J Exp Psychol* 4(1):11–26
- Ho C, Spence C (2005) Olfactory facilitation of dual-task performance. *Neurosci Lett* 389(1):35–40
- Hosking S, Young K, Regan M (2007) The effects of text messaging on young novice driver performance. In: Faulks IJ, Regan M, Stevenson M, Brown J, Porter A, Irwin JD (eds) *Distracted driving*. Australasian College of Road Safety, Sydney, pp 155–187
- Hyman R (1953) Stimulus information as a determinant of reaction time. *J Exp Psychol* 45:188–196
- Just MA, Keller TA, Cynkar J (2008) A decrease in brain activation associated with driving when listening to someone speak. *Brain Res* 1205: 70–80, Elsevier
- Laberge-Nadeau C, Maag U, Bellavance F, Lapierre SD, Desjardins D, Messier S, Saïdi A (2003) Wireless telephones and the risk of road crashes. *Accid Anal Prev* 35(5):649–660
- Laoufi M (2005) Development of a comprehensive vibration model for evaluation of rumble strips. In: Eskandarian A *Masters thesis*, The George Washington University, Washington, DC
- Laux LF, Mayer DL (1993) Informational aspects of vehicle design. In: Peacock B, Karwowski W (eds) *Automotive ergonomics*. Taylor and Francis, London, pp 401–429
- Lee JD (2006) Driving safety. Reviews of human factors and ergonomics. Human Factors and Ergonomics Society, Santa Monica
- Lee JD et al (2009) Defining driver distraction, Chapter 3. In: Regan MA, Lee JD, Young KL (eds) *Driver distraction: theory, effect and mitigation*. CRC Press, Boca Raton, pp 31–40
- MacAdam CC (2003) Understanding and modeling the human driver. *Veh Syst Dyn* 40, Nos. 1–3:101–134

- Madden M, Lenhart A (2009) Teens and distracted driving-texting, talking and other uses of the cell phone behind the wheel. Pew Research center. <http://pewinternet.org/Reports/2009/Teens-and-Distracted-Driving.aspx> Washington, DC, 16 Nov 2009
- Mallis M, Maislin G, Konowal N, Byrne V, Bierman D, Davis R et al (2000) Biobehavioral responses to drowsy driving alarms and alerting stimuli. Final report to develop, test and evaluate a drowsy driver detection and warning system for commercial motor vehicle drivers sponsored by the National Highway Traffic Safety Administration, Federal Highway Administration, Office of Motor Carriers
- McKnight AJ, Adams BB (1970) Driver education ad task analysis. Volume I task descriptions, final report, contract no. FH 11-7336. Human Resource Research Organization, Alexandria
- McKnight AJ, Adams BB (1970) Driver education ad task analysis. Volume II task analysis methods, final report, contract no. FH 11-7336. Human Resource Research Organization, Alexandria
- Mehmood A, Easa MS (2009) Modeling reaction time in car following behavior based on human factors. World academy of science, engineering and technology issue 57, conference, Sept 2009, Article 122, pp 710. <http://www.waset.org/journals/waset/v57/v57-122.pdf>
- Michon JA (1985) A critical view of driver behavior models: what do we know, what should we do? In: Evans L, Shwing RC (eds) Human behavior and traffic safety. Plenum Press, New York, pp 485–520
- Michon JA (ed) (1993) Generic intelligent driver support. Taylor and Francis, London
- Mironer M, Hendricks D (1994) Examination of single vehicle roadway departure crashes and potential IVHS countermeasures. DOT HS 808 144, Aug 1994
- Nakayasu H, Nakagawa M, Miyoshi T, Abe H (2010) Human cognitive reliability analysis on driver by driving simulator. In: IEEE 40th international conference on computers and industrial engineering (CIE). Awaji, pp 1–6
- Neale VL, Dingus TA, Klauer SG, Sudweeks J, Goodman M (2005) An overview of the 100 car naturalistic study and findings. In: Proceedings of the 19th international technical conference on the enhanced safety of vehicles. National Highway Traffic Safety Administration, Washington, DC
- Olson PL, Sivak M (1986) Perception-response time to unexpected roadway hazards. Hum Factors 28(1):91–96
- Olson PL, Dewar R, Farber E (2010) Forensic aspects of driver perception and response, 3rd edn. Lawyers & judges, Tucson
- Ozguner U, Stiller C, Redmill K (2007) Systems for safety and autonomous behavior in cars: the DARPA grand challenge experience. In: Proceedings of the IEEE, advanced automotive technologies, Feb 2007, 95(2): 397–412
- Peacock B, Karwowski W (eds) (1993) Automotive ergonomics. Taylor and Francis, London
- Pickrell T, Ye T (2008) Driver electronic device use in 2008, NHTSA traffic safety facts, DOT HS 811 184. Washington, DC. (<http://www-nrd.nhtsa.dot.gov/pubs/811184.pdf>)
- Ranney T (2008) Driver distraction: a review of the current state-of-knowledge. National Highway Safety Administration, Washington, DC
- Reed N, Robbins R (2008) The effect of text messaging on driver behavior: a simulator study. Transport Research Laboratory, Berkshire
- Sigari MH (2009) Driver hypo-vigilance detection based on eyelid behavior, advances in pattern recognition, ICAPR '09. In: Seventh international conference on advances in pattern recognition, 4–6 Feb 2009, pp 426–429
- Sommer D, Golz M, Trutschel U, Edwards D (2009) Assessing driver's hypovigilance from biosignals. In: Proceedings of the 4th European conference of the international federation for medical and biological engineering, Springer, Berlin, pp 152–155
- Spence C, Ho C (2009) Cross modal information processing in driving. In: Castro C (ed) Human factors of visual and cognitive performance in driving. CRC Press, Boca Raton, pp 187–196
- Stevens A, Parks A (2001) ADVISERS-a strategic approach to ADAS development. In: IEEE international conference on advanced driver assistance systems (ADAS), 17–18 Sep 2001, vol 483, pp 1–3
- Strayer DL, Drews FA, Johnston WA (2003) Cell phone-induced failures of visual attention during simulated driving. J Exp Psychol Appl 9(1):23–32
- Strayer DL, Drews FL, Crouch DJ (2006) A comparison of the cell phone driver and the drunk driver. Hum Factors 48(2):381–391

- Strayer DL, Drews FL (2004) Profiles in driver distraction: effects of cell phone conversations on younger and older drivers. *Hum Factor: J Hum Factor Ergon Soc* 46:640–649
- Tapani NL (2009) Human reaction times as a response to delays in control systems – notes in vehicular context. Kajaani Unit of Department of Information Processing Science, University of Oulu. V20091106. <http://www.measurepolis.fi/alma/ALMA%20Human%20Reaction%20Times%20as%20a%20Response%20to%20Delays%20in%20Control%20Systems.pdf>
- TRAFFIC SAFETY FACTS 2009, DOT HS 811 402, National Highway Traffic Safety Administration, National Center for Statistics and Analysis
- U.S. Department of Transportation 2009 <http://www.nrd.nhtsa.dot.gov/Pubs/811402EE.pdf>
- Vahidi A, Eskandarian A (2003) Research advances in intelligent collision avoidance and adaptive cruise control. *IEEE Trans Intell Transp Syst* 4(3):143–153
- Wang J, Knippling RR (1994) Single vehicle roadway departure crashes: problem size assessment and statistical description. DOT HS 808 113
- Wilson FR, Sinclair JA, Bisson BG (1989) Evaluation of driver/vehicle accident reaction times. University of New Brunswick, Fredrickson, The transportation Group
- Wierwille WW (1993) Visual and manual demands of in-car displays. In: Peacock B, Karwowski W (eds) *Automotive ergonomics*. Taylor and Francis, London, pp 299–319
- Wierwille WW (1999) Historical perspective on slow eyelid closure: whence PERCLOS?. In: *Ocular measures of driver alertness*, technical conference proceedings. Herndon, pp 130–143
- Web Sites**
- (Audiworld Website) <http://www.audiworld.com/news/07/audi-q7-42-tdi/content.shtml>
- (Autonews Website) <http://autonews.gasgoo.com/china-news/kia-s-luxury-cadenza-sedan-to-sell-in-china-by-jun-100429.shtml>
- (Daimler website) <http://media.daimler.com/dcmedia/0-921-614216-1-1147529-1-0-0-0-0-0-11702-0-0-1-0-0-0-0-0.html>
- (Expert Review Website) <http://www.expertreviews.co.uk/car-tech/1283890/new-ford-focus>
- (Ford Website) http://media.ford.com/pdf/Drive_smart_futuretec.pdf
- (Honda Website) http://world.honda.com/news/2003/4030618_2.html
- (Inficon Website) [http://www.inficon.com/technology/avoid-blind-spots-while-driving-by-using-the-proper-mirror-adjustments/196/\(blind-spot-detection\)](http://www.inficon.com/technology/avoid-blind-spots-while-driving-by-using-the-proper-mirror-adjustments/196/(blind-spot-detection))
- (Ivsource website) http://ivsource.net/archivep/2001/feb/010212_nissandemo.html
- (Japanvehicles Website) <http://www.japanvehicles.com/newcars/toyota/Caldina/main.htm>
- <http://localdc.com/alphardhybrid.htm>
- (psa-peugeot-citroen Website) <http://www.sustainability.psa-peugeot-citroen.com/corporate-citizenship/priorities/overview/achievement.htm?id=2708>
- (wikipedia Buick Website) <http://en.wikipedia.org/wiki/Buick>
- (wikipedia Website) http://en.wikipedia.org/wiki/Lane_departure_warning_system
- (worldcarfans Website) <http://www.worldcarfans.com/10503019431/the-new-honda-legend-acura-rl>
- Web Site: “DARPA Grand Challenge”; http://en.wikipedia.org/wiki/DARPA_Grand_Challenge
- Web site: US DoT Driving Distraction: <http://www.distraction.gov/stats-and-facts/index.html>
- Web site: Onstar System: <http://www.onstar.com/web/portal/emergencyexplore?tab=1>
- Web site (Freie Universitat Berlin- Artificial Intelligence Lab, AutoNOMOS Labs, Prof. Rojas): <http://autonomos.inf.fu-berlin.de/>

20 Driver Behavior Modeling

Samer Hamdar

The George Washington University, Ashburn, VA, USA

1	<i>Introduction</i>	538
2	<i>Driver Behavior Modeling Framework</i>	539
3	<i>Operational Stage Acceleration Models</i>	544
3.1	Fundamental Diagram	548
3.2	Trajectories	549
4	<i>Tactical Stage Lane Changing Models</i>	550
5	<i>Conclusion</i>	556

Abstract: In this chapter, the author presents a general framework classifying the different models adopted for capturing driver behavior focusing on the human cognitive dimensions and the traffic decision-making dimensions. Special interest is directed toward the “lower-level” microscopic models that can be linked directly to two core driving assistance technologies: adaptive cruise controls and lane-departure warning systems. These “lower-level” models are classified either as acceleration models or as lane changing models.

Acceleration models are at the core of operational driving behaviors, and include car-following models which capture interactions between a lead vehicle and following vehicles. The main assumption in these models is that the behavior of the following vehicle is directly related to a stimulus observed/perceived by the driver, defined relative to the lead vehicle. In addition to the operational aspect, lane changing models capture the tactical side of driving. Most lane changing models have followed a deterministic rule-based framework where changing lanes is directly related to the desirability of such maneuver, its necessity, and its possibility/safety. Recognizing the limitations of the major existing microscopic traffic models, the objective in this chapter is to advance the state of knowledge in modeling driver behavioral processes and to offer an insight into current modeling approaches and the corresponding advantages and disadvantages.

1 Introduction

Since the 1950s, traffic scientists have tried understanding driver behavior through the construction of different traffic models (Chandler et al. 1958). The main focus is on recreating the different congestion dynamics observed on “our roads” while keeping the models simple enough so that proper calibration and validation can be conducted. Traffic scientists have adopted different modeling approaches trying to realize higher fidelity and robustness in capturing the different traffic regimes. From a simulation perspective, three approaches have been suggested: the macroscopic approach, the mesoscopic approach, and the microscopic approach. The macroscopic approach considers traffic as a flow bounded by the geometric characteristics of the road along which it moves. The flow is characterized mainly by three macroscopic measures of effectiveness: density (number of vehicles per unit distance of a roadway at a given time), flow rate (number of vehicles passing a given location per unit time), and average speed (aggregate speed measures during a time period at a given location, that is, time-mean speed and space-mean speed). The vehicles cannot be distinguished from each other and trajectory data cannot be obtained as the flow moves following given density-flow-speed relationships. The microscopic approach tries imitating traffic phenomena as a collective manifestation of inter-vehicle interactions. Basic elementary rules are used while vehicles are represented explicitly in the simulation and trajectories are retraced along the study time period. Finally the mesoscopic approach is a hybrid approach representing vehicles individually as in the microscopic approach but using macroscopic density-flow-speed relationships

to move them. Since this chapter is dedicated to understanding individual driver behavior, the author focuses on microscopic traffic modeling.

On the other hand, microscopic models can be based on concepts taken from different disciplines ranging from psychology to physics. There is no definitive superiority of one type of the models over the others especially that each approach has its advantages and its disadvantages and that different model logics have different execution-time horizons. While focusing on the shortest execution-time horizons (acceleration and lane changing), this chapter is structured as follows: the first section defines the different families of driving behavioral models from a traffic science perspective and links such definition to the cognitive human decision-making framework adopted by psychologists. The second section presents the major acceleration models and assesses their ability of capturing congestion dynamics. The core lane changing models are offered in [Sect. 3](#). [Section 4](#) summarizes the main findings of the background review.

2 Driver Behavior Modeling Framework

Microscopic traffic models capture driver behavior at the individual decision level, collectively giving rise to aggregate traffic flows. Extensive research is conducted with the objective to capture phenomena associated with congestion dynamics (such as flow breakdown on freeways and hysteresis) and incident scenarios. However, challenges are being faced especially when studying the safety dimension: most existing driving decision models are built in a “crash-free” environment with no cognitive dimensions and with a limited ability to describe driver behavior under extreme and incident conditions. The behavioral maneuvers that allow for a safer driving environment are not completely understood. One of the challenges is then to use an existing driving decision framework and link it to a cognitive decision framework that can support different psychological and physiological states (i.e., hysteresis-related behavior, panic, etc.).

Driver decisions can be viewed as taking place at five different levels with associated time horizons (FHWA 2004):

1. Pre-trip: decisions made before starting a trip (departure time choice, mode choice).
2. Strategic en-route: decisions that drivers make en-route, while executing a trip. These decisions usually impact the overall structure of the trip (route choice and switching).
3. Tactical route: a consequence of “small multi-part decisions that are made in order to complete a small but coordinated portion of a trip” (lane changing, overtaking).
4. Operational driving: decisions that a driver makes “on a near instantaneous basis to satisfy an immediate goal” (acceleration, gap acceptance).
5. Vehicle control: decisions made instantaneously to “satisfy human-machine interaction needs.”

These five levels are shown in [Table 20.1](#) in relation to the approximate time needed to execute the decisions at each level.

■ Table 20.1
Classification of Driver Behavioral Models

Driver Behavioral Model	Time to Make and Execute Decision
Pre-Trip	>1 hour
Strategic En Route	30/60 seconds – 1 hour
Tactical Route Execution*	5 seconds – 30/60 seconds
Operational Driving*	Instantaneous – 5 seconds
Vehicle Control	Mechanical/Electric Specifications Related

*Focus of this Chapter

Looking into the human cognitive aspect of driving and how it can be related to the vehicle control, the main focus of this chapter is on the tactical route execution decisions and the operational driving decisions a driver makes when traveling on freeways and uninterrupted highway facilities: such decisions are key in the current development of the intelli-drive systems. The major model categories developed to capture tactical route execution decisions at the individual passenger car level for freeway traffic are the lane changing models, the merging models, the passing and overtaking models, and the cooperative behavior models. The major model categories developed to capture the operational driving decision processes consist of the acceleration models, the gap acceptance models, and the queue discharge models. At this stage, the main focus is on the acceleration models category and the lane changing models category, for which a detailed review is presented in the next two sections.

One of the challenges in capturing cognitive processes while keeping the framework presented earlier is the number of factors that influence driving behavior and the complexities that arise when trying to incorporate them in the corresponding models. Ranney (1999) summarized the main factors incorporated in existing car-following models; these factors are:

1. Time headway or distance headway
2. Relative velocity
3. Degree to which the following vehicle is tracking the velocity changes of the leading vehicle
4. Stream speed
5. Driving goals
6. Length of time vehicle is in a given driving state
7. Whether car-following is elected or imposed
8. Road curvature

The influence of the first three factors has been reported empirically, while limited empirical studies have been performed to test the significance of the other five factors (Rothery 1999; Boer 1999; Treiber et al. 2000).

For a more complete analysis, the different factors affecting car-following behavior can be divided into two main categories: individual differences and situational factors. The former are “permanent” attributes of the driver and vehicle, such as:

1. Age
2. Gender
3. Wealth (income level, type of job)
4. Education
5. Race
6. Physiological characteristics:
 - (a) Reaction time
 - (b) Body strength (strength of legs, strength of grip, muscle coordination, presence of a disability)
 - (c) Vision
7. Family characteristics (single, children or not, single child, all male or female siblings)
8. Impatience level
9. Aggressiveness or risk-taking propensity
10. Skills (includes motor skills and cognitive/decision-making skills, anticipation, memory, learning)
11. Vehicle characteristics:
 - (a) Vehicle size
 - (b) Vehicle performance characteristics:
 - (i) Strength (engine power)
 - (ii) Maneuverability (sports car, SUV, sedan, truck)
 - (iii) Visibility factor
 - (iv) Grip of tires
 - (v) Car age

The situational factors vary over time and space, and depend on the traffic conditions surrounding a given driver. They can be either related to the individual driver or the environment in which the vehicle is driven. The main situational environmental factors are:

1. Weather (rain, sun, ice, snow, cold)
2. Visibility (sun, dust, rain)
3. Noise level
4. Day of week
5. Time of day
6. Network characteristics:
 - (a) Road characteristics and their dynamics:
 - (i) Asphalt conditions and asphalt type
 - (ii) Road type (weaving, rural)
 - (iii) Number of lanes

- (iv) Access Type (freeway, highway, arterial, ramp, secondary road)
 - (v) Grade
- (b) Traffic laws and their dynamics:
 - (i) Speed limit
 - (ii) Presence of cameras/police cars
 - (iii) Traffic signs (stop signs, yield)
 - (iv) Traffic markings
 - (v) Traffic signals
 - (vi) Presence of surrounding work zone or incident (signs or no signs)
- (c) Traffic characteristics and their dynamics
 - (i) Congestion level: lateral and longitudinal spacing with surrounding vehicles (location)
 - (ii) Speed and acceleration of surrounding vehicles (first and second moments of location)
 - (iii) Type of surrounding vehicles

The main situational individual factors are:


1. Distraction
2. Impairment (alcohol, drugs, stress, fatigue, etc.)
3. Trip characteristics and their dynamics:
 - (a) Trip purpose (reach destination or enjoy scenery or both)
 - (i) Type of activity at destination
 - (ii) Type of activity at origin
 - (b) Length of drive:
 - (i) Time of activity at destination (desired or forecasted arrival time)
 - (ii) Time of activity at origin (departure time choice)

The above list of factors and their classification is not unique and is based on a synthesis performed by the author on the basis of several types of literature encountered in different disciplines. The first type consists of media flyers with information derived from driver questionnaires recorded in census traffic data (National Highway Traffic Safety Administration, U.S. Dept. of Transportation 1998). The second type of literature consists of research reports assessing different traffic models with empirical validation (Ranney 1999; FHWA 2004; Hamdar and Mahmassani 2008). Finally, the third type of literature is derived from social and psychological research studies (Schultz 1964; Helbing et al. 2000; Schmidt and Warner 2002; Wallsten et al. 2005) adapted by the author to the traffic context. For example, when studying the disrespect of laws under different conditions (normal versus extreme conditions) (Schultz 1964), it is reported that when the purpose is to save someone's life (trip purpose: evacuation with high urgency), a person tends to disrespect rules conflicting with such purpose (traffic laws, signs) according to his/her personality (aggressiveness versus conservative), possibly contributing to a system failure or collapse (breakdown of the transportation system/network).

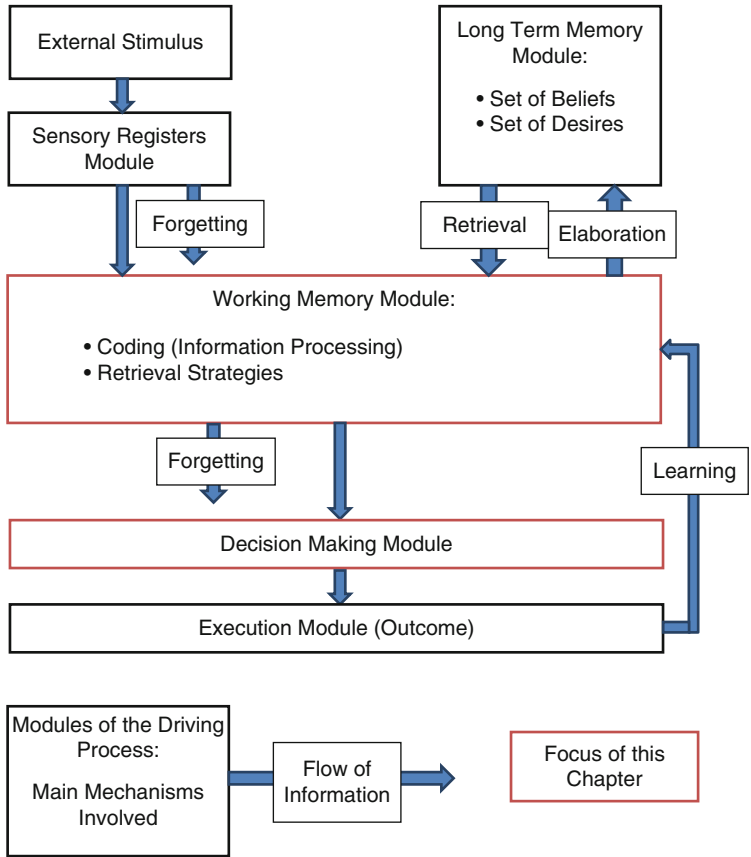
Some overlap and possible high correlation may be present among some of the listed factors. For example, when it is raining, the grip of the tires on the road surface and visibility will decrease. The exact interrelation between the factors and how they influence drivers is not well understood. This fact, combined with the high number of factors to be incorporated, may have caused traffic modelers to rely on simplifying assumptions in representing these behaviors. The main assumptions adopted in the literature are:

1. Drivers respect safe following rules so that no accidents can be created.
2. The following distance is based on the time required to perceive the need to decelerate and react by applying the brakes.
3. The car-following response (acceleration or deceleration) is a function of a stimulus represented by the velocity difference with respect to the lead vehicle and a sensitivity term.
4. The most successful stimulus in terms of explaining car-following behavior is the relative speed divided by the spacing, while the least successful stimulus is the spacing between vehicles (Ranney 1999).
5. Drivers in the car-following mode mainly respond to the vehicle in front of them. Some of the models also include look-ahead, whereby drivers may use information from more than one vehicle ahead to predict the behavior of the lead vehicle (Brackston and McDonald 1999).

Since there is no single model taking into consideration all the exogenous factors mentioned earlier, the above assumptions reduce the number of parameters in existing traffic simulation models and help simplify the calibration/validation task. However, such models rarely incorporate cognitive dimensions such as risk-taking and risk-perception behaviors. The challenge faced by researchers is to use a sufficient number of parameters to capture cognitive processes (perception, judgment, and execution) while keeping the model practical in terms of implementation and calibration effort.

To identify the significant influencing parameters and incorporate them in a suitable microscopic model, better understanding of the cognitive decision-making processes of drivers is helpful. Accordingly, the author refers to the human information-processing system introduced by Hastie and Dawes (2001) to the driving task, shown in  Fig. 20.1. In this framework, the working memory module and the decision-making modules are fundamental to the problem especially for devising intelli-drive systems logic. The external stimulus module, the sensory registers module, and the execution module can be directly related to different audiovisual sensors and mechanical/electrical devices which a vehicle can be equipped with.

In summary, acceleration operational models and lane changing tactical models constitute the main subject of study in this chapter. The author is interested in the soft side of the problem that is the logic adopted for information coding (perception) and decision making (judgment).



■ Fig. 20.1
Information-processing/execution system for the driving activity

In the next two sections, the author presents related research findings in the area of modeling acceleration and lane-changing behaviors. This exposition allows a better appreciation of the framework presented earlier. A comparative assessment of several major acceleration models is conducted.

3 Operational Stage Acceleration Models

In previous continuous-time single-lane car-following models, the main response to a given stimulus is performed through acceleration or deceleration (stimulus-response or General Motors Models – Gazis et al. 1961). The stimulus consists of the velocity of

the driver, the relative velocity between a vehicle and the front vehicle, and finally, the corresponding space-gap (Gazis et al. 1961). One limitation of these models is their inapplicability under very low traffic densities. Another is that in dense traffic, small gaps will not induce braking reactions if the front vehicle is traveling at the same velocity (zero relative velocity).

Newell (1961) addressed this problem by introducing the concept of the velocity depending adiabatically on the gap. Like all previous car-following models, the Newell model is collision-free. Moreover, since there is an immediate dependence of the velocity on the density (gap), very high and unrealistic accelerations can be produced. To overcome this limitation, Bando et al. (1995) modified Newell's model by controlling the change in velocity by a relaxation time, resulting in the Optimal Velocity Model (OVM). The model is known to produce possibly unrealistic accelerations when the relaxation time is less than 0.9 s.

The so-called Generalized Force Model (Tilch and Helbing 1998) proposed a generalized optimal velocity function that incorporates reaction to velocity differences and different rules for acceleration and braking. However, even though this model was able to produce time-dependant gaps and velocities, unrealistic small accelerations and decelerations were produced as well. While the above models (Newell, OVM and Generalized Force Model) offer important insight into the car-following logic, they are not included in the comparative detailed assessment presented next, due in part to the known issues they face and also to the fact that the character of the behavior they produce are subsumed in other models. The focus is on the following models:

1. Gazis, Herman, and Potts' (GHP) model (Gazis et al. 1959)
2. Gipps' model (Gipps 1981)
3. Cellular automaton Model (CA) (Nagel and Shreckenberg 1992)
4. S-K Model by S. Krauss (Krauss and Wagner 1997)
5. IDM model or Intelligent Driver Model (Treiber et al. 2000)
6. IDMM model or Intelligent Driver Model with Memory (Treiber et al. 2006)
7. Wiedemann Model (Wiedemann and Reiter 1992)

Each of these seven models' formulation is presented first and then assessed especially in terms of fundamental flow-density diagrams and trajectory data.

GHP Model. The GHP family of models is a standard stimulus-response model that links the stimulus (relative speed) to the response (acceleration) by a sensitivity parameter. The relation that governs the GHP Models is:

$$a_{n+1}(t + \tau) = \lambda z[v_n(t) - v_{n+1}(t)] \quad (20.1)$$

where

τ = reaction time

$a_n(t)$ = acceleration of a vehicle n at time t

$v_{n+1}(t)$ = position of a vehicle $n + 1$ following vehicle n at time t
 λz = sensitivity term

The sensitivity term λz has received most of the attention in earlier research (Gazis et al. 1959, 1961). It took different functional forms as described below:

1. $\lambda z = c$, that is, a constant.
2. $\lambda z = \begin{cases} w & \text{if } s \leq s_{critical} \\ w' & \text{if } s > s_{critical} \end{cases}$, that is, a step function. s is the spacing ($y_{n+1} - y_n - l_{n+1}$) between two vehicles. l_n is the length of vehicle n and y_n is its position. $s_{critical}$ is a threshold specified by the modeler.
3. $\lambda z = \frac{c}{s}$. This form is adopted in the GHP model and is called reciprocal spacing. c is a constant.
4. $\lambda s = \frac{c \cdot \dot{y}_{n+1}}{s^2}$ where \dot{y}_{n+1} is a speed term. This form is adopted in Edie's Model (Edie 1961).
5. $\lambda z = \frac{c}{s^2}$ leads to the famous macroscopic Greenshield's flow-density relation.

Having the spacing s in the denominator $\lambda z = \frac{c}{s}$ above will reduce the acceleration response (increase deceleration rates) tremendously for smaller headways. Moreover, assuming that the driver will be able to observe and measure exactly the relative speed term, the vehicle will travel at the same speed as the leader when $[v_n(t) - v_{n+1}(t)] = 0$.

Gipps Model. The Gipps Model is characterized by two modes of driving. The first mode is a free-flow mode where a vehicle is trying to reach its desired velocity based on a given acceleration pattern. The second mode is a car-following mode where the velocity of the driver is the "safe velocity" allowing him/her to avoid rear-end collision. Safe velocity is dictated by a maximum deceleration rate tolerated by the driver. The driver chooses the minimum between the velocity obtained from the car-following mode and the velocity obtained from the free-flow mode:

$$v_n(t + \tau_n) = \min \left\{ \begin{aligned} &v_n(t) + 2.5a_n\tau_n \left(1 - \frac{v_n(t)}{v_{0n}} \right) \left(0.025 + \frac{v_n(t)}{v_{0n}} \right)^{1/2}; \\ &b_n\tau_n + \sqrt{b_n^2\tau_n^2 - b_n \left[2(y_{n-1}(t) - l_{n-1} - y_n(t)) - v_n(t)\tau_n - \frac{v_n(t)^2}{b_{n-1}} + D_n \right]} \end{aligned} \right. \quad (20.2)$$

where:

$y_n(t)$ = position of the front bumper of vehicle n (following a vehicle $n-1$ at time t)
 $v_n(t)$ = velocity of vehicle n at time t
 l_n = effective length of vehicle n
 b_n = desired deceleration rate of vehicle n
 a_n = desired acceleration rate of vehicle n
 τ_n = reaction time of vehicle n

CA Model. The continuous limit of the CA model (Krauss et al. 1996) is defined as follows:

$$\begin{aligned} v_0 &= \min[v(t) + a_{\max}, v_{\max}, s_{\text{gap}}(t)], \\ v(t+1) &= \max[0, v_{\text{des}} - b_{\max} n_{\text{ran},0,1}], \\ y(t+1) &= y(t) + v(t+1) \end{aligned} \quad (20.3)$$

where:

$s_{\text{gap}}(t)$ = the free space to the vehicle ahead

a_{\max} = maximum acceleration

$n_{\text{ran},0,1}$ = random number in the interval (0, 1)

b_{\max} = maximum deceleration due to the noise

This rule-based logic allows the driver to choose between a maximum speed, a safe speed $s_{\text{gap}}(t)$, and a free-flow speed $v(t) + a_{\max}$. Some randomness due to the b_{\max} parameter is introduced.

S-K Model. The S-K Model follows a logic similar to the CA Model's logic:

$$\begin{aligned} v_1 &= \min[v(t) + a_{\max}, v_{\max}, v_{\text{safe}}], \\ v_2 &= v_1 - \varepsilon l \{v_1 - [v(t) - b_{\max}]\}, \\ v(t+1) &= v_{\text{ran},v_2,v_1}, \\ x(t+1) &= y(t) + v(t+1) \end{aligned} \quad (20.4)$$

where

$a_{\max} = b_{\max}$ = acceleration/deceleration rate corresponding to a given vehicle

v_{ran,v_2,v_1} = random term between the “optimal” velocity v_1 and the deviation from that velocity v_2

εl = parameter determining the deviation from the optimal velocity

v_{safe} = safe velocity

The main difference between the CA Model and the S-K Model is in the term v_{safe} (velocity below which no accidents are generated). The CA model sets v_{safe} to the gap between two consecutive vehicles s_{gap} and thus producing unrealistic deceleration rates. The S-K Model sets v_{safe} based on a maximum allowable deceleration rate following the same logic adopted in the Gipps Model.

IDM/IDMM Model. The IDM and IDMM models assume that the acceleration is a continuous function of the velocity v_n , the gap s_n , and the velocity difference Δv_n :

$$\dot{v}_n = a_{\max}^{(n)} \left[1 - \left(\frac{v_n}{v_0^{(n)}} \right)^{\delta l} - \left(\frac{s^*(v_n, \Delta v_n)}{s_n} \right)^2 \right] \quad (20.5)$$

This expression can be seen as the integration of two tendencies. The first tendency is to accelerate with:

$$a_f(v_n) = a_{\max}^{(n)} \left[1 - \left(\frac{v_n}{v_0^{(n)}} \right)^{\delta l} \right] \quad (20.6)$$

The second tendency is a deceleration tendency: when vehicle n comes too close to the leading vehicle, drivers tend to brake with a deceleration of:

$$b_{\text{int}}(s_n, v_n, \Delta v_n) = -a_{\max}^{(n)} \left(\frac{s^*(v_n, \Delta v_n)}{s_n} \right)^2 \quad (20.7)$$

The desired gap s^* is set by the following equation:

$$s^*(v, \Delta v) = s_0^{(n)} + s_1^{(n)} \sqrt{\frac{v}{v_0^{(n)}}} + TS^{(n)} v + \frac{v \Delta v}{2 \sqrt{a_{\max}^{(n)} b^{(n)}}} \quad (20.8)$$

where $s_0^{(n)}$, $s_1^{(n)}$, $v_0^{(n)}$, $TS^{(n)}$, $a^{(n)}$, and $b^{(n)}$ are parameters that may be calibrated for each vehicle n .

Wiedemann Model. Based on Wiedemann's work (Wiedemann and Reiter 1992), a driver is assumed to have four driving modes:

1. Free-Flow Mode: a vehicle is not influenced by any front vehicle
2. Approaching Mode: a driver consciously influenced because of his/her perception of a slower vehicle in front of him/her
3. Following Mode: a driver unconsciously following a vehicle
4. Emergency Situation: when the headway between consecutive vehicles drops below a desired "safety distance"

These modes are determined by two main factors: the speed difference (relative velocity) and the distance between two successive vehicles (space headway).

3.1 Fundamental Diagram

Based on the initial description of each model, the flow-density relationships (i.e., fundamental diagrams) can be plotted and assessed. The GHP model, the original Gipps model, and the S-K Model were not able to capture the metastable congested state nor the instability encountered at the beginning of the traffic breakdown (Hamdar and Mahmassani 2008); the metastable state corresponds to the maximum flow state that appears just before congestion and breakdown to congestion. This state is characterized by formation of a small horizontal plateau near capacity (reported to be between 1,800 and 2,200 veh/h.lane) with some deviation from the linear relation between flow and density (Tadaki et al. 2002). The absence of the metastable state was already mentioned in several publications in the literature in the case of the GHP model and the Gipps model (FHWA 2004, Treiber et al. 2006). However, even though the S-K Model is a simplified

version of the Gipps model, previous studies indicated that the S-K Model offers the advantage of capturing congested traffic behavior due to the deceleration randomization inherited from the CA model (Krauss and Wagner 1997). This behavior is characterized by the observation of multiphase states in the fundamental diagram (congestion buildup, stop and go waves, then traffic deterioration).

On the other hand, the CA model is reported to produce traffic breakdowns when reaching a flow capacity of 1,800 veh/h. However, the CA lacks the cognitive logic behind it, making the model somewhat mechanical and sometimes, unrealistic. This is due to the fact that the model is controlled heavily by the constant deceleration rate attributed to the drivers. As for the IDM model, although still more improvement is needed on the cognitive dimension, this model has improved on the CA model in that respect. Both IDM and IDMM showed realistic fundamental diagrams with a stable region and an unstable region.

Finally, the Wiedemann model showed the same congestion instability exhibited by the IDM and the IDMM models. Moreover, its complexity, though a disadvantage for some researchers, allows a more realistic and complete view of the different factors encountered in the driving task.

3.2 Trajectories

After reporting the basic fundamental diagrams' assessment, a closer microscopic look is presented in this subsection. Based on trajectory time-location data reported for each vehicle, the GHP models allow vehicles to follow each other at high speeds with extremely small space headway. Another unrealistic behavior is observed in the CA Model; as suggested in the literature, a vehicle can follow a leader with almost zero meters separating it from the lead vehicle: it is forced to stop at that location using an unrealistic deceleration rate (safety constraint). The Gipps Model and the S-K Model show vehicles moving in platoons: drivers are allowed to travel closely to each other because they consider applying a maximum deceleration rate at all instances to avoid accidents (safe velocity concept described in the Gipps model and the S-K Model descriptions). The two models that reportedly allowed more uniform and larger space headways are the same models that captured traffic instability during congestion: IDM Model, IDMM Model, and Wiedemann Model.

In conclusion, over the years, acceleration models were improved in terms of capturing congestion dynamics and traffic disturbances. The latest models (Wiedemann, CA, IDM, IDMM) produced realistic fundamental diagrams and trajectories. However, no single model of the presented models is characterized with enough simplicity for calibration/validation purposes, with a cognitive dimension for better driver decision-making understanding, and with suitable stability to create realistic crashes when relaxing the safety constraints.

Acceleration models capture only the operational aspect of drivers' behavior (🔗 Table 20.1). The tactical aspect is captured mainly by lane changing models; the main existing lane changing models are presented in the next section.

4 Tactical Stage Lane Changing Models

Starting with the Gipps Model (1986), lane changing models are mostly rule-based models where drivers are assumed to follow a series of if-then rules structured into different decision-making levels. This modeling approach is adopted by several authors with further improvement to account for forced merging and courtesy lane changing (Hidas 2002). Wiedemann and Reiter (1992) introduced perceptual thresholds in the lane changing decision-making logic acknowledging that drivers perceive and respond to stimulus in an imperfect manner. Finally, random utility theory was adopted in the lane changing process (Ahmed 1999). This brought a new perspective to the problem but with a considerable increase in the required calibration effort. The main lane changing models are described next.

Gipps Model (1986). The decision-making process in the Gipps' Model is based on three questions:

1. Is it possible to change lanes?
2. Is it necessary to change lanes?
3. Is it desirable to change lanes?

The main assumption is that drivers have two objectives. The first objective is to keep a given desired speed. The second objective is to use a feasible lane ("correct lane") allowing exiting a network section in a safe and comfortable manner (through an intersection or an interchange). These two objectives may be conflicting with each other and realizing them depends on the following set of influencing factors (Gipps 1986):

1. The physical possibility of changing lanes
2. The presence of a permanent obstruction and its location
3. Lane type
4. The driver's intended turning movement
5. The presence of heavy vehicles
6. Speed

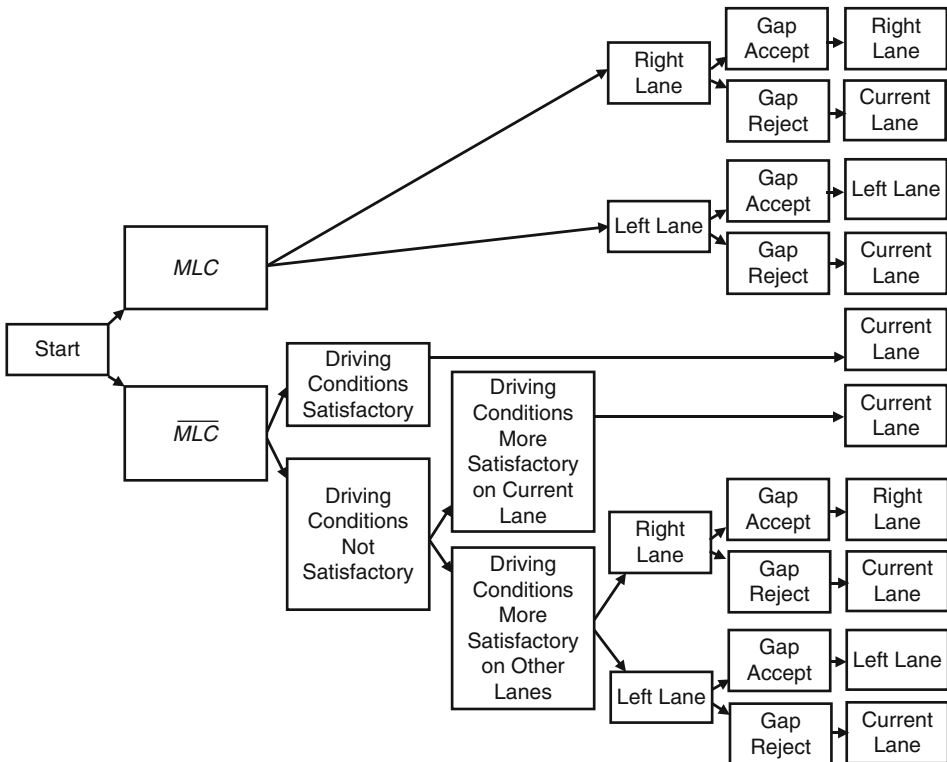
Drivers can follow one of three patterns of behavior based on the relative distance to the intended turn (Factor 4 above). If the driver is considered at a remote distance of the intended turn, his/her main concern is to maintain the desired speed. When the turn is at an intermediate distance, drivers start considering moving toward the intended turn, and lane changing opportunities to improve the vehicle's speed will start to be ignored. The final driving pattern is met when the driver is close to the turn and the interest is solely limited to be in the correct lane. Despite the ease of implementing this model, the rule-based framework imposes a deterministic structure for the tactical driver decision-making process. Safety rules as imposed by the Gipps car-following model are still imposed here while no inter- and intra-driver heterogeneity is captured.

Ahmed Model (1999). In this model, lane changing is divided into two classes: Mandatory Lane Changes (MLC) and Discretionary Lane Changing (DLC). It is assumed that the decision process adopted by the drivers is based on:

1. Whether to consider lane changing or not
2. Which lane should be targeted
3. Whether to accept the gap in the chosen target lane

Moreover, drivers are assumed to make lane changing decisions (DLC) at discrete points in time. These decisions are independent of previous lane changing decisions and maneuvers. However, in the forced merging (MLC), the impact of previously made decisions will be taken into account. The main behavioral components of the model are presented below.

The Lane Changing Process. The lane changing decision process is latent in nature. The only observable part is the completion of the execution of the lane change. Representing the observable parts by rectangles and the latent parts by ovals, [Fig. 20.2](#) presents the conceptual framework of the lane changing process. The main idea is that drivers will change lanes if they decide to respond to lane changing encouraging conditions whether they are mandatory lane changing conditions or discretionary lane changing conditions. The way the driver will respond is modeled using the random utility approach (Ben-Akiva and Lerman 1985). Panel data is used to estimate the corresponding model.



■ Fig. 20.2

MITSIM's lane changing framework

If the driver sees that the driving conditions on the current lane are not satisfactory, these conditions will be compared to driving conditions on other adjacent lanes.

On the other hand, the effort due to changing lanes will be captured by positively correlated utilities of the adjacent two lanes. Due to the nested-type decision, the nested logit model structure (Ben-Akiva and Lerman 1985) is adopted: first, the utility of the two adjacent lanes are compared. Then, the utility of the current lane is compared to the respective utilities of the adjacent lanes to determine the lane to be chosen. Once the above process is performed, the probability of selecting each of the lanes (current, left, and right lanes) will be obtained. If one of the adjacent lanes is selected, drivers will search for an acceptable gap.

The Gap Acceptance Process. The main assumption is that drivers consider only the adjacent gap, that is, the gap between the lead vehicle and the lag vehicle on the target lane. This gap is composed of the lead gap and the lag gap. The minimum acceptable lead and lag gap lengths are the critical lead and lag gaps, respectively. They vary under different traffic situations. The critical gap for driver n at time t has the following form:

$$G_n^{cr,g} = \exp(X_n^g(t)\beta a^g + \alpha^g v_n + \varepsilon a_n^g(t)) \quad (20.9)$$

where:

$$g \in \{lead, lag\}$$

α^g = parameter of speed of v_n for $g \in \{lead, lag\}$

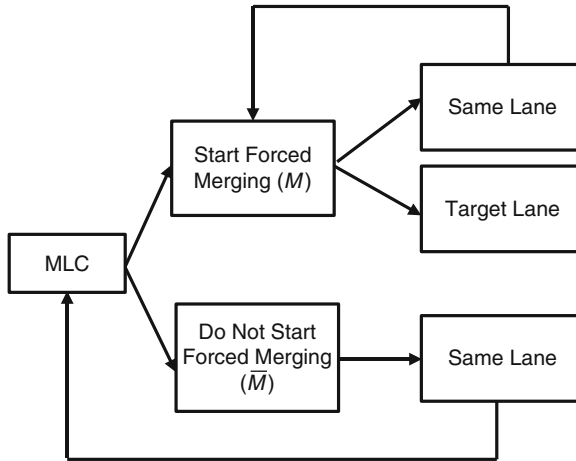
$\varepsilon a_n^g(t)$ = generic random term that varies across the three dimensions g , t , and n

v_n and α^g capture the correlation between the lead and the lag critical gaps, for a driver. This correlation is expected to be positive. If $\varepsilon a_n^g(t)$ is assumed to be normally distributed with $N(0, \sigma_{\varepsilon}^2)$, the critical gaps are log-normally distributed and the conditional probability of accepting a gap will be determined accordingly.

The Forced Merging Behavior. The tree diagram illustrating the forced merging behavior logic differs from that presented in [Fig. 20.2](#). The corresponding structure is shown in [Fig. 20.3](#) where the ovals still represent the latent parts involving decisions and the rectangles represent the observable events.

At each time step, a driver will evaluate the traffic conditions in the target lane so a decision can be reached on whether the driver will try to merge in front of the lag vehicle in the target lane. Accordingly, he/she will try to communicate with the lag vehicle to see if the right of way is established or not. When the right of way is established, forced merging (denoted by M) will start. This may take less than a second up to a few seconds. However, if the right of way is not established, the driver will continue the “evaluation/communication” process in the next time step.

Using a binary logit model with a suitable random utility specification to the corresponding panel data, the conditional probability of switching from state M to state



■ Fig. 20.3
MITSIM's forced merging framework

M is determined. The main exploratory variables considered in the forced merging are (Ahmed 1999):

1. Lead relative speed only when the lead vehicle is slower
2. Lag relative speed
3. Remaining distance to the point at which lane changing must be completed by
4. Delay (time elapsed since the mandatory lane change conditions apply)
5. Total clear gap (the sum of the lead and lag gaps)
6. Indicator for heavy vehicles

The main advantage of the above model is its probabilistic nature while considering several correlation and heterogeneity measures. However, this requires more effort in the calibration process.

Wiedemann and Reiter Model (1992). As in the Gipps Model, before deciding to change lanes, a driver needs to answer the following three questions:

1. Desire question: Is there a desire to change lanes?
2. Condition question: Is the present driving situation at neighboring lane favorable?
3. Possibility question: Is the movement to a neighboring lane possible?

Lane changes are then classified into two major groups: lane changes to faster lanes and lane changes to slower lanes. Lane changing to faster lanes is the result of an obstruction on the actual lane: this obstruction is caused by a slower vehicle not allowing the vehicle of interest to reach a given desired speed. Lane changing to a slower lane is due to the general tendency to keep right or to the need to move out of the way of a faster vehicle (tailgating). This lane changing type is only accepted if there are no vehicles obstructing the target lane within a certain time limit.

Both lane changing types are only accepted if the lane changing maneuver is considered safe. This is checked by taking into account the distance headways and the relative speeds with respect to lead and lag vehicles on the target lanes. The modelers rely on the distance threshold used in the Wiedemann acceleration “psychophysical model” to define six categories of lane changing maneuvers. These categories are explained in [Table 20.2](#).

Even though it takes into consideration perceptual uncertainties before changing lanes, this model is deterministic in nature. Moreover, it requires an extensive calibration effort due to the large number of parameters involved.

Hidas Model ([Hidas 2002](#)). Implemented in SYTRAS software, the “Hidas” lane changing logic is called at each update interval. The components of this logic explained next.

- 1. Is lane changing necessary? The lane changing maneuver can be considered necessary for different reasons evaluated sequentially in their order of importance. The factors considered in the evaluation process are ([Hidas 2002](#)):
 - (a) Turning Movement
 - (b) End-of-lane
 - (c) Incident (Lane Blockage)
 - (d) Transit Lane
 - (e) Speed Advantage
 - (f) Queue Advantage

■ **Table 20.2**
Types of lane changes in the Wiedemann Model

Lane changes to faster lanes	
Free lane changes	The vehicle of interest is only influenced by the leading vehicle on the current lane. The leading and the lagging vehicle on the target lane are not considered in the maneuver
Lead lane changes	The leading vehicle on the target lane is closer to the vehicle of interest than the leading vehicle on the current lane and the lagging vehicle on the target lane is not influenced by the maneuver
Lag lane changes	The lagging vehicle of the target lane is influenced by the maneuver and the leading vehicle on the current lane is closer than the leading vehicle on the target lane
Gap lane changes	The lagging vehicle on the target lane is influenced by the maneuver and the leading vehicle on the target lane influences the vehicle of interest
Lane changes to slower lanes	
Free lane changes	The maneuver is not influenced by the lagging vehicle on the target lane
Acceleration lane changes	The lagging vehicle on the current lane influences the vehicle of interest

2. Selection of target lane. The choice set for the target lane is limited to lane(s) adjacent to the current lane. Some factors causing lane changes may determine the direction of such maneuver (to the left or to the right). These factors include the location of a given exit/entrance and the presence of transit lanes. Other factors will have no effect on the direction of the lane change (incident, queue advantage, and speed) where drivers choose the target lane depending on the corresponding traffic and safety conditions.
3. Is lane changing to the target lane feasible? A lane change is feasible if there is a sufficiently large gap in the target lane so that the subject vehicle can move safely without forcing other vehicles on the target lane to slow down significantly. For this purpose, the potential leader and follower in the target lane are found. The lane change is feasible if:
 - (a) The deceleration or acceleration required for the subject vehicle to move behind the lead vehicle on the target lane is acceptable.
 - (b) The deceleration required for the follower in the target lane to allow the subject vehicle to move into the lane is acceptable.
4. Driver courtesy in the target lane. In SITRAS, forced lane changing is based on a courtesy level offered by drivers in the target lane. In other words, when a driver wishes to change lanes, he/she will send a courtesy request to followers in the target lane; these drivers will evaluate this request and will accept it or reject it based on different factors such as speed, position, and driver type. When the “courtesy” is sent by a driver, he/she will decelerate to ensure the creation of a large enough gap in the next few seconds.
5. Change of the target lane. In the Hidas Model, the lane change is not modeled explicitly. Once the lane changing decision is feasible, the vehicle will be inserted in the new lane after a given time period (lane changing time).

The SYTRAS lane changing logic is one of the few logics capturing forced and cooperative lane changing maneuvers; however, at its base, this model could be viewed as an extension of a rule-based Gipps Model. Moreover, due to the different modules included in the lane changing logic, this model is computationally demanding without allowing anticipation and risk-taking behavior.

MOBIL Model (Kesting et al. 2007). MOBIL discusses the rationale behind changing lanes from an operational standpoint. The main assumption is that during the lane changing decision process, a driver makes a compromise between the expected improvements on his/her driving conditions and the disadvantages imposed on other drivers, mainly the followers on the target lane.

The advantages and the disadvantages of changing lanes, or the utility of changing lanes, are expressed in terms of difference in accelerations after and before changing lanes. Such utility formulation allows for:

1. The acceleration function in the car-following model to be used in assessing the safety and the desirability of changing lanes and thus, for a more compact and concise model.
2. The acceleration model and the lane changing model to be consistent with each other: any anticipation or additional complexity involved in the acceleration model is directly transferred to the lane changing model.

3. Safety conditions considered in the acceleration model to be kept: the braking deceleration imposed on the new follower in the target lane should consider safety rules and is an element of the motivation for changing lanes.
4. General use of the model in conjunction with several acceleration models.

Since MOBIL Lane Changing Model deals only with the operational level of the decision process, the criteria to change lanes are divided into motivational criteria and safety criteria. In the motivational criterion, the main influence parameter is expressed in terms of a “politeness parameter” that will take into account the discomfort caused by a driver on his surrounding when changing lanes. The motivation for changing lanes can be purely “egoistic” (politeness factor of 0) or “altruistic” (politeness factor of 1). When the politeness parameter allows a lane change only when the combined accelerations of the lane-changing driver and his neighbors increase, the resulting strategy is called “Minimizing Overall Braking Induced by Lane Changes” and thus the acronym MOBIL.

Even though it can be applied in conjunction with different car-following models, the MOBIL model has several weak points:

1. Myopic approach ignoring anticipation and anisotropy. However, this may be the compromise made to retain a concise and simple model.
2. Some conditions/equations are problematic, especially when generalizing the model to a more complete network. For example, limiting the choice of lanes based on the higher motivational incentive (difference in acceleration) is too restrictive.
3. MOBIL is difficult to integrate with some acceleration models (models that are velocity based rather than acceleration based – example the Gipps Model).
4. It may be argued that the acceleration decision while lane changing is different than the acceleration decision while car-following.
5. This model is only applied with the IDM acceleration model at this stage, and has undergone little calibration effort.

5 Conclusion

After presenting the driving behavioral modeling framework in [Sect. 1](#), this chapter presents a detailed literature review on the major operational stage (acceleration) and tactical stage (lane changing) traffic models. The comparative assessment of the major acceleration models in [Sect. 2](#) reveals that there is no one model that can capture driver behavior in free-flow, congested, and accident-prone regimes. Three main models allow capturing congestion dynamics (IDM, IDMM, and Wiedemann) but are not designed to exhibit stability when safety constraints are relaxed. The more stable Wiedemann Model allows the incorporation of some cognitive dimension but at the expense of a larger number of parameters to be estimated (13 parameters). The review of lane changing models in [Sect. 3](#) reveals that most of these models are rule-based models. Recent improvements to these models have enhanced their behavioral content and practical

realism, for example, by incorporating forced merging behavior, distinguishing between mandatory and discretionary lane changing, and introducing politeness factors. However, most reviewed models remain deterministic and somewhat myopic in terms of decision-making processes with no possibility of accidents or of accounting for perception errors.

References

- Ahmed KI (1999) Modeling driver's acceleration and lane changing behaviors. Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51:1035–1042
- Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis. MIT Press, Cambridge MA
- Boer ER (1999) Car following from the driver's perspective. *Transportation Research Board, Part F* 2(4):201–206
- Brackston M, McDonald M (1999) Car-following: a historical review. *Transportation Research, Part F* 2(4):181–196
- Chandler R, Herman R, Montroll W (1958) Traffic dynamics: studies in car-following. *Oper Res* 6:165–184
- Eddie LC (1961) Car-following and steady-state theory for non-congested traffic. *Oper Res* 9:66–76.
- FHWA (2004) NGSIM Task E.1-1: core algorithms assessment, final report, Cambridge Systematic, Inc., Massachusetts
- Gazis DC, Herman R, Potts R (1959) Car-following theory of steady state traffic flow. *Oper Res* 7:499–505
- Gazis D, Herman R, Rothery R (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Gipps PG (1981) A behavioral car-following model for computer simulation. *Transportation Research* 15B:101–115
- Gipps PG (1986) A model for the structure of lane changing decisions. *Transportation Research* 20B:403–414
- Hamdar SH, Mahmassani HS (2008) From existing accident-free car-following models to colliding vehicles: exploration and assessment, National Research Council (US). *Transportation Research Board Meeting*, Washington, DC, 87th January 2008. Preprint CD-ROM
- Hastie R, Dawes RM (2001) Rational choice in an uncertain world. Sage, Thousand Oaks
- Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 406:487–491
- Hidas P (2002) Modeling lane-changing and merging in microscopic traffic simulation. *Transportation Research, Part C* 10(2):351–371
- Kesting A, Treiber M, Helbing D (2007) MOBIL: general lane changing model for car-following models. *Transportation Research Record* 1999/2007:86–94
- Krauss S, Wagner P (1997) Metastable states in a microscopic model of traffic flow. *Phys Rev E* 55(5):5597–5602
- Krauss S, Wagner P, Gawron C (1996) Continuous limit of Nagel-Shrekenberg model. *Phys Rev E* 54(4):3707–3712
- Nagel K, Shrekenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I (France)* 2:2221–2229
- National Highway Traffic Safety (1998) Aggressive driving; help get the world out, US. DOT. <http://purl.access.gpo.gov/GPO/LPS3277>. Accessed December, 2004
- Newell G (1961) Nonlinear effects in the dynamics of car-following. *Oper Res* 9:209–229
- Ranney TA (1999) Psychological factors that influence car-following and car-following model development. *Transportation Research, Part F* 2(4): 213–219
- Rothery RW (1999) Traffic flow theory: A state-of-the-Art report-revised monograph on traffic flow theory. *Transportation Research Board, National Research Council*, Washington, DC
- Schmidt LJ, Warner B (2002) Panic: origins, insight, and treatment. North Atlantic Books, California
- Schultz DP (1964) Panic behavior, discussion and readings. Random House, New York

- Tadaki S, Nishinari K, Kikuchi M, Sugiyama Y, Yukawa S (2002) Analysis of congested flow at the upper stream of a tunnel. *Physica A* 315: 156–162
- Tilch B, Helbing D (1998) Generalized force model of traffic dynamics. *Phys Rev E* 58:133
- Treiber M, Hennecke K, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 2(2):1805–1824
- Treiber M, Kesting A, Helbing D (2006) Delays, inaccuracies and anticipation in microscopic traffic models. *Physica A* 360:71–88
- Wallsten TS, Pleskac TJ, Lejuez CW (2005) Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychol Rev* 112(4):862–880
- Wiedemann R, Reiter U, (1992) Microscopic traffic simulation, the simulation system mission. Project ICARUS (V1052) Final Report, CEC, Brussels

21 Using Naturalistic Driving Research to Design, Test and Evaluate Driver Assistance Systems

Gregory M. Fitch¹ · Richard J. Hanowski²

¹Virginia Tech Transportation Institute-Truck and Bus Safety,
Blacksburg, VA, USA

²Center for Truck and Bus Safety, Virginia Tech Transportation
Research Institute, Blacksburg, VA, USA

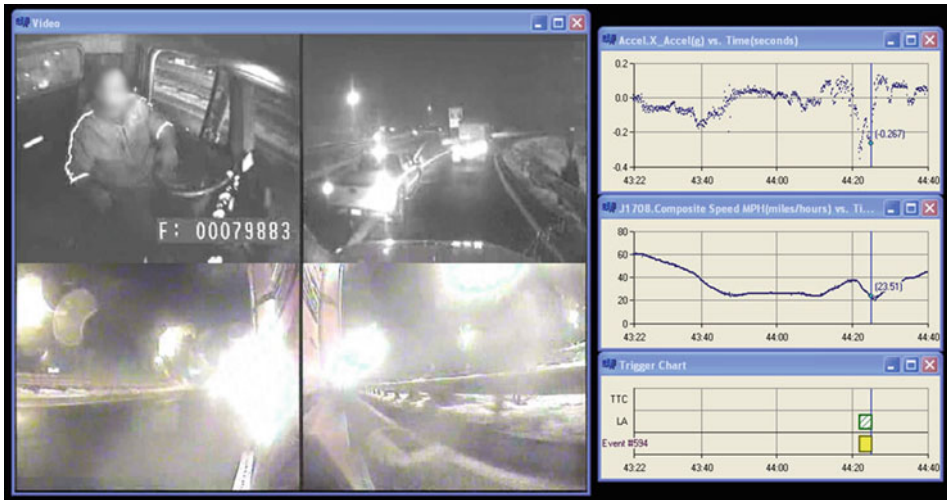
1	<i>What Is Naturalistic Driving Research?</i>	560
2	<i>Naturalistic Driving Research: Bridging Epidemiology and Empirical Research</i>	562
3	<i>NDS Methods</i>	563
3.1	Data Acquisition System (DAS)	563
3.2	Vehicle Instrumentation	565
3.3	Vehicle Data Collection	565
3.4	Participant Data Collection	566
3.5	NDS Data Reduction	566
3.6	NDS Data Analysis	569
4	<i>Using NDSs in Designing Driver Assistance Systems</i>	571
4.1	Using NDSs to Identify User Requirements	571
4.2	Using NDSs to Guide the Design of Driver Assistance Systems	572
4.3	Using NDSs to Test Driver Assistance Systems	573
4.3.1	Testing Early-Stage Prototypes	574
4.3.2	Testing Working Prototypes	574
4.4	Using NDSs to Evaluate Driver Assistance Systems	575
5	<i>Conclusion</i>	577
6	<i>Case Study</i>	577

Abstract: Naturalistic driving research is the in situ investigation of driver performance and behavior. Video cameras and a suite of sensors are installed on participants own vehicles and are used to continuously record the driver, the vehicle, and the environment over an extended period of time. The collected data typically span hundreds of thousands of vehicle-miles-traveled and provide an “instant replay” of the rare occurrence of safety-critical events. The method supports the representative design of experiments, where the drivers, vehicles, and environment sampled are representative of the conditions to which the results are applied. Naturalistic Driving Studies (NDS) are an effective tool for the design, testing, and evaluation of driver assistance systems. This is because they can support various stages of a user-center systems design process. First, NDSs can help determine what drivers need from a new driver assistance system by allowing researchers to assess the driver error contributing to safety-critical events. Secondly, the approach can serve the testing of working prototypes, where “natural” driver behavior and performance with the candidate driver assistance systems is observed. Thirdly, novel test criteria, such as drivers’ rate of involvement in safety-critical events, can be used to evaluate the driver assistance systems’ effectiveness at improving driver performance. NDSs and their role in the design, testing, and evaluation of driver assistance systems are described in this chapter.

1 What Is Naturalistic Driving Research?

Naturalistic driving research is the in situ investigation of driver performance and behavior. That is, driver performance and behavior is observed as it occurs in the full context of real-world driving. Naturalistic driving research consists of installing video cameras and a suite of sensors on participants own vehicles and continuously recording the driver, the vehicle, and the environment over an extended period of time (e.g., 4 weeks, 1 year). Because participants are not given any instructions and there is little interaction between them and the experimenter, “natural” driving behavior and performance is observed. The produced datasets typically span hundreds of thousands vehicle-miles-traveled, facilitating the “instant replay” of the seconds leading up to the rare occurrence of a safety-critical event (🔗 [Fig. 21.1](#), for example, shows a safety-critical event in which a driver rapidly decelerated because he was cut off by a lead vehicle). Furthermore, scientific hypotheses can be tested using the representative design of experiments (Hammond and Stewart 2001). This is because the drivers, vehicles, and environment sampled in the dataset are representative of the conditions to which the results are applied. Because driver performance and behavior are heavily affected by the environment, NDSs can provide interesting insight on how drivers’ adapt to changing environmental conditions.

Naturalistic driving research, using instrumented vehicles, was pioneered at the Virginia Tech Transportation Institute (VTTI) in 1984 (Dingus 2008). 🔗 [Figure 21.2](#) shows the first test vehicle instrumented with a video camera at VTTI. The method quickly evolved between 1989 and 2005, in part from the miniaturization of hardware



■ Fig. 21.1

Screen capture of a safety-critical event recorded in a recent naturalistic driving study of truck driver performance with camera/video imaging systems (Fitch et al. [in press](#)). Driver's face has been obscured to protect privacy



■ Fig. 21.2

An instrumented vehicle at VTTI in 1984 (Dingus [2008](#))

and the sophistication of software. However, its evolution was also made possible because the National Highway and Traffic Safety Administration (NHTSA) and the Federal Highway Administration (FHWA) recognized that NDSs were the most effective and accurate means of obtaining meaningful data relevant to the safety issues that the administrations were addressing (Dingus [2008](#)). These transportation administrations were particularly interested in unobtrusively collecting driving data that would allow the characterization of real-world driver performance and their interaction with in-vehicle

technologies. To date, the NDS method has been used to collect driver performance and behavior data from truck drivers, passenger vehicle drivers (with specific studies performed using teenage drivers and elderly drivers), as well as motorcycle riders.

2 Naturalistic Driving Research: Bridging Epidemiology and Empirical Research

Transportation safety has been traditionally investigated using epidemiological and empirical research. Epidemiology is a retrospective research approach whereby crash databases, for example, are analyzed to investigate crash characteristics. This approach is particularly effective for assessing non-behavioral components of the crash including environmental (e.g., weather), infrastructure (e.g., number of road lanes), and vehicle (e.g., brake failure) conditions. As highlighted in the landmark “Indiana Tri-Level Study” (Treat et al. 1979), the environmental, infrastructure, and vehicle factors are three key areas pertaining to crashes that can be well-studied using epidemiological methods. However, the “human factor” associated with crashes, which is suggested to be the most critical aspect of most crashes, is more difficult to assess through epidemiology.

Why is this the case? Crash databases used in epidemiological research are populated by Police Crash Reports (often called police accident reports, PARs), which make it difficult (if not impossible) to determine with certainty driver behavior immediately preceding the crash. For example, questions such as “where did the driver look in the seconds preceding the crash?” cannot be ascertained through a PAR (and if that information was included in the PAR, how reliable would it be? For instance, drivers have trouble remembering, or can refuse to admit, what they did prior to the crash). Given the circumstances under which PAR data are collected, it is not possible to reliably determine driver behaviors that may have contributed to the crash. As such, determining cause-and-effect relationships are not possible through the epidemiological approach. Rather, epidemiological studies tend to focus on assessing the increase in risk given certain circumstances. For example, determining that not wearing a safety-belt significantly increases the likelihood of a fatality in a crash.

To be able to make cause-and-effect statements on driver behavior and performance, empirical research is the most appropriate method. Laboratories, driving simulators, test tracks, and the open road are frequently used because of the internal validity gained by being able to carefully control conditions. Typically, participants engage in test scenarios where behavior and performance are measured. For example, empirical studies using the Virginia Smart Road test track have used naïve truck drivers to test and evaluate systems that reduce blind-spots on tractor-trailers (Wierwille et al. 2007). Empirical studies such as this provide important data to examine benefits and unintended consequences of safety technologies by studying driver behavior and performance when interacting with these systems. At the same time, it can be difficult and costly to systematically control all driver, vehicle, and environmental factors. As such, empirical research becomes limited in the degree to which findings generalize to real-world conditions.

NDSs serve as a hybrid between epidemiological and empirical methods. By observing drivers in their own vehicles without an experimenter present, everyday behavior and performance is evinced as participants quickly forget they are being recorded. Furthermore, the reasons for making a trip pertain to the participant, not the researcher. Yet, when a safety-critical event is recorded, detailed information on where the driver looked, what the driver did, what the vehicle dynamics were, and what the environmental conditions were in the seconds leading up to the event can be meticulously analyzed. As such, NDSs can be thought of as an answer to Egon Brunswik's call in the 1940s for Psychology to give equal consideration to organisms and the environment through the representative design of experiments (Hammond and Stewart 2001). Representative design refers to the arrangement of conditions in the experiment such that they represent the conditions to which the results are applied. The systematic execution of representative design research (i.e., the combination of empirical and representative design research) was heavily criticized in the 1950s because of the labor, time, and costs involved. However, as exhibited by NDSs, technological advancements have made this approach feasible, facilitating the investigation of how drivers behave in changing environmental conditions.

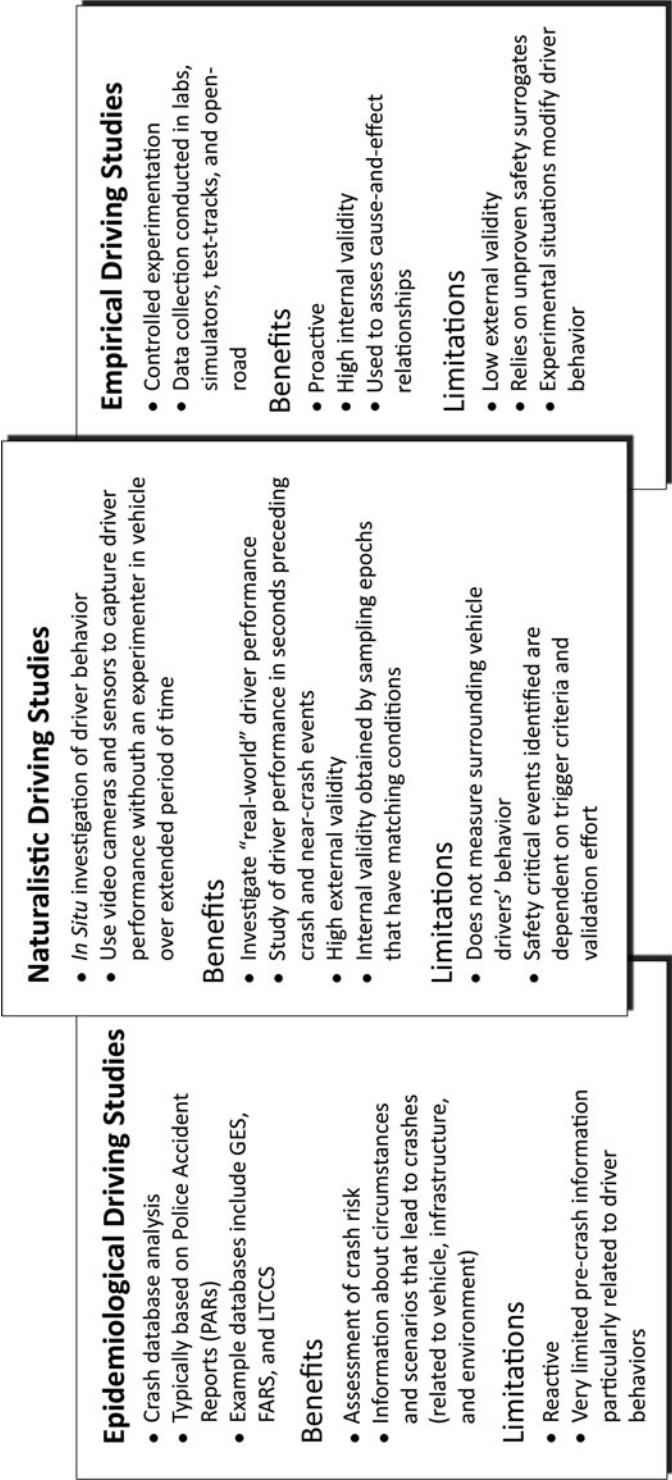
Adapted from (Dingus 2003), ● Fig. 21.3 shows the relationship between epidemiological, empirical, and naturalistic research. It also shows that the various research methods used to assess driver behavior have their own strengths and limitations. It is important to note that each method discovers essential information pertaining to traffic safety.

3 NDS Methods

An overview of the methods used to perform an NDS is provided below. The data acquisition system, data collection procedure, and data reduction procedure are briefly described.

3.1 Data Acquisition System (DAS)

The DAS is the central component to any NDS. The main computing unit receives and stores data from a network of sensors distributed throughout the vehicle. The sensors include an interface box to the vehicle network, an accelerometer for longitudinal and lateral acceleration, a radar system that measures range and range rate to lead and trailing vehicles in the adjacent lanes, a video-based lane-tracking system that measures lane-keeping performance, a light meter, a temperature/humidity sensor, a GPS sensor for location, and continuous video recordings of the driver and roadway to validate the sensor-based findings. Data are stored on an external hard drive, which can store several weeks of driving data before requiring replacement. The DAS automatically starts when the vehicle's ignition is turned on, and automatically powers down when the ignition is turned off. Data are recorded directly to a hard drive that is locked in place to prevent tampering. Processor speed, data storage capacity, system robustness, and reliability are



■ Fig. 21.3 Highlights, benefits, and limitations of three data collection approaches used to evaluate driver behavior and performance (Adapted from Dingus (2003))



■ Fig. 21.4

CAD illustration of the next generation DAS developed by VTTI (Klauer et al. [in press](#))

crucial for continuously collecting naturalistic driving data. The DAS units used in earlier NDSs were mounted in inconspicuous locations. In the commercial motor vehicle studies, they were mounted under the passenger seat (Blanco et al. [in press](#); Hanowski et al. 2008), while in the light vehicle studies they were mounted in the vehicle's trunk (Dingus et al. 2006). Recently, a next generation DAS was developed to be smaller than a VHS tape and mount inside the vehicle cabin under the rear-view mirror (🔗 [Fig. 21.4](#)).

3.2 Vehicle Instrumentation

Depending on proximity and availability, test vehicles are instrumented either at the test site, or at the research institute. The DASs are typically installed in a few hours, with additional time being required to install sensors unique to the NDS being performed. Quality control checks are performed by driving the vehicle for a few hours and inspecting each sensor's operation.

3.3 Vehicle Data Collection

Vehicle data collection involves a technician traveling to the vehicles on a weekly basis to inspect the test equipment for proper operation and to retrieve the collected data. Test vehicles are located using their GPS coordinates. If any equipment is damaged, the technician makes the necessary repairs on site. During these visits, the technician also replaces the hard drives with empty hard drives to ensure there is sufficient storage for the subsequent data collection interval. The technician has access to the hard drive as to not require the participant to be present.

The collected data are then downloaded once the hard drives return to the research institute. A random sample of ten files from each hard drive is inspected for data issues. A quality assurance report is generated and circulated to the data collection team so that the appropriate repairs can be scheduled. Each file on the hard drive is then inspected to verify that the appropriate driver was operating the test vehicle. This is done because non-participating drivers sometimes operate the test vehicle. These files are deleted upon their discovery because the drivers did not consent to have their data included in the dataset.

3.4 Participant Data Collection

All participants are met with at the onset of the study to read and sign an informed consent form. Demographic, as well as physiological data (e.g., visual acuity), are typically collected at this time. Some NDSs have involved administering a questionnaire to drivers on a bi-weekly basis throughout the study, while others have only administered a post-study questionnaire. In a recent NDS study of driver performance with a Camera/Video Imaging System (C/VISs), drivers completed a questionnaire pertaining to their use of the C/VIS, which helped identify how their driving performance changed over time.

3.5 NDS Data Reduction

Once the data has been ingested onto a secure server at the research institute, it is ready to be processed. To identify events of interest in the dataset, researchers write algorithms that scan the dataset and create pointers to potential events. These pointers were termed “triggers” at VTTI because this approach is used to trigger potential safety-critical events, such as crashes and near-crashes. Examples of simplistic logic used in previous NDSs to identify safety-critical events are shown in [Table 21.1](#).

It should be noted that the lower the trigger criteria are set, the more false-positive triggers, non-conflict triggers, and less severe conflicts are identified. The trade-off is that lower trigger values result in relatively few missed valid events. The goal of setting a lower, optimum trigger value is to identify all of the most severe events (crashes and near-crashes) without having an unmanageable number of false-positive triggers, non-conflict triggers, and low-severity conflict events.

► [Figure 21.5](#) shows an example of a valid trigger for longitudinal acceleration (i.e., a braking event). In this example, the Trigger Chart shows the trigger at the same point that the Accel_X (longitudinal acceleration) plot shows the value reached -0.267 g, indicating a sharp deceleration of the vehicle. For this example, the longitudinal acceleration trigger was set at -0.20 g; therefore, anytime the software detected a longitudinal acceleration with a magnitude greater than -0.20 g, a trigger was created. Looking closely at the video in the top right quadrant, a pickup truck can be seen in front (and to the left) of the instrumented vehicle. At this point, the pickup truck begins to

■ Table 21.1
Triggers and trigger values used to identify safety-critical events

Trigger type	Definition	Description
Longitudinal acceleration (LA)	Hard braking or sudden deceleration	Deceleration greater than or equal to 0.20 g. Speed greater than or equal to 1.6 km/h
Time-to-collision (TTC)	The amount of time (in s) that it would take for two vehicles to collide if one vehicle did not perform an evasive maneuver	A forward TTC value of less than or equal to 2 s, coupled with a range of less than or equal to 76 m, a target speed of greater than or equal to 8 km/h, a yaw rate of less than or equal to $ 6^{\circ}/s $, and an azimuth of less than or equal to $ 0.12^{\circ} $
Swerve (S)	A sudden “jerk” of the steering wheel to return the truck to its original position in the lane	Yaw greater than or equal to 2 rad/s^2 . Speed greater than or equal to 8.05 km/h
Analyst identified (AI)	An event that is identified by the analyst but has not been identified by a trigger	Event that was identified by a data analyst viewing video footage; no other trigger listed above identified the event (i.e., LA, TTC, etc.)



■ Fig. 21.5
Example of a validated trigger where the LA was of a greater magnitude than the threshold of -0.20 g (Fitch et al. [in press](#))

change lanes, crossing the solid lane line directly into the lane in front of the instrumented vehicle. The driver of the instrumented truck brakes to avoid contact with this pickup truck.

Each triggered event is inspected by a team of data reductionists. Their role is to: (1) visually verify that the triggered event is indeed valid, (2) answer a series of questions pertaining to the event (e.g., how severe was the event, what was the traffic density at the time, was the driver talking on a cell phone, etc.), and (3) perform a frame-by-frame reduction of where the driver looked in the seconds centering around the event. The answers to the questions that reductions provide comprise the study’s data directory. A standard data directory was developed by Hickman et al. (in press) when performing a field operational test of a drowsy driver warning system. The same data directory has also been used in Blanco et al. (in press). By standardizing the data directories, NDS datasets can be combined to increase sample size for future studies. For instance, a combined dataset was used in an investigation of commercial motor vehicle driver distraction (Olson et al. 2009).

A primary question in the data directory regards the severity of the safety-critical event. Table 21.2 shows the criteria used by reductionists to classify a valid safety-critical event’s severity.

Table 21.2
Description of safety-critical event severity categories

Event type	Description
Crash	Any contact with an object, either moving or fixed, at any speed
Curb strike: avoidable	Any contact with a curb or median where it is apparent that the driver could have performed a maneuver to avoid the contact
Curb strike: unavoidable	Any contact with a curb or median where it is apparent that the driver could not have performed a maneuver to avoid the contact. The most likely cause for these events is poor roadway design
Near-crash	Any circumstance that requires a rapid, evasive maneuver (e.g., hard braking, steering) by the subject vehicle (SV) or any other vehicle, pedestrian, cyclist, or animal, in order to avoid a crash
Crash-relevant conflict	Any circumstance that requires a crash-avoidance response on the part of the SV, any other vehicle, pedestrian, cyclist, or animal that was less severe than a rapid evasive maneuver (as defined above), but greater in severity than a normal maneuver. A crash-avoidance response can include braking, steering, accelerating, or any combination of control inputs
Illegal maneuver	Any circumstance where either the SV or the other vehicle performs an illegal maneuver such as passing another vehicle across the double yellow line or on a shoulder. For many of these cases, neither driver performs an evasive action

However, triggered events do not necessarily need to be safety-critical events. Algorithms can be used to trigger other driving maneuvers, such as left turns and successfully executed lane changes. The verification process also does not need to be performed (although, it helps ensure a high level of validity). Sampling triggered events can help refine algorithms such that the sampling errors are tolerable. This, however, requires substantial consideration, as a researcher needs to ensure that the data is what they think it is before conducting analyses.

Triggers are also used to sample baseline driving epochs. Baseline epochs are instances where no safety-critical event takes place. The capture of baselines is critically important when investigating the risk of specific behaviors. This is because they allow researchers to compare the prevalence of factors in safety-critical events to the prevalence of the same factors when safety is not jeopardized. To sample baseline epochs, baseline triggers are programmed to randomly sample instances in the dataset based on a driver's exposure (e.g., vehicle-miles-traveled, hours of operation, or involvement in safety-critical events). A team of reductionists then validates each baseline trigger and applies a standard data dictionary. In the previously mentioned investigation of commercial motor vehicle driver distraction, a total of 19,888 baseline epochs were sampled to assess the risk of specific non-driving activities (Olson et al. 2009).

3.6 NDS Data Analysis

There are multiple approaches to analyzing NDS data. An important analysis is the quantification of the risk of involvement in a safety-critical event when a specific condition is present. As mentioned above, odds ratios have been computed to investigate the risk of specific mobile device subtasks while driving. These analyses revealed that complex subtasks, such as texting and dialing on a cell phone, increase risk, while conversing on a cell phone, does not increase risk. In fact, in certain conditions, it was found to decrease risk for both light vehicle and commercial motor vehicle drivers. ▶ [Table 21.3](#) summarizes the odds ratio results produced from three NDSs.

NDSs also afford the investigation of specific driving maneuvers. For example, the amount of time drivers take to execute a lane change has been assessed using light vehicle and commercial motor vehicle NDSs (Fitch et al. [in press, 2009](#)). Furthermore, how close drivers get to an adjacent vehicle when changing lanes has been assessed using rear-facing radars. This is the first time that clearance data has been captured during a commercial motor vehicle NDS and provides interesting insight on these maneuvers.

Another important analysis is the investigation of drivers' visual behavior when executing maneuvers. The video data collected of the driver's face can be inspected to determine where the driver looked on a frame-by-frame basis. This is done without the use of invasive eye-tracking equipment attached to the driver. For example, such investigations have revealed where drivers look prior to a safety-critical event, prior to changing lanes, and when executing left-hand turns. These analyses have assessed the probability of a driver looking in a specific location, the number of glances made to a specific location,

■ **Table 21.3**
Odds ratios for mobile device use reported in naturalistic driving studies of CMV and light vehicle drivers

		Naturalistic driving study		
		Hickman et al. (2010)	Olson et al. (2009)	Klauer et al. (2006)
Vehicle type examined		Tractor-trailers, three-axle trucks, and transit buses	Tractor-trailers	Light vehicles
Data collection interval		9/08 – 9/09	5/04 – 5/05 and 11/05 – 5/07	1/03 – 7/04
Number of vehicles		13,305	205	109
Number of crashes		1,085	21	69
Total number of SCEs		40,121	4,452	830
Number of baselines		211,171	19,888	20,000
Task	Cell phone use (collapsed)	OR = 1.22 ^a	OR = 1.04	Not computed
	Reaching for headset/earpiece	OR = 3.4 ^a	OR = 6.72 ^a	Not computed
	Reaching for cell phone	OR = 3.8 ^a	Included in dialing phone	Not computed
	Texting/e-mailing/accessing the internet	OR = 163.6 ^a	OR = 23.24 ^a	Not computed
	Dialing cell phone	OR = 3.1 ^a	OR = 5.93 ^a	OR = 2.8 ^a
	Talking/listening hands-free cell phone	OR = 1.31	OR = 0.44 ^a	Not observed
	Talking/listening hands-held cell phone	OR = 0.78 ^a	OR = 1.04	OR = 1.3

^aindicates a significant OR. A significant OR greater than one indicates an increased risk, while a significant OR less than one indicates a decreased risk

and the mean duration spent looking at a specific location. Scanning patterns can also be assessed with this data, such as the probability of transitioning from one location to another (Wierwille 1981).

It is well known that drivers’ brake response time is affected by the environmental conditions that are present. A strength of NDS datasets is that they can be used to assess drivers’ Brake Response Time (BRT) in rear-end conflicts. Although the goal is not to create one canonical BRT, BRTs for various conditions can be estimated by controlling the conditions selected from the dataset.

Overall, NDSs offer a novel perspective from epidemiological and empirical research. Specific aspects of driver performance and behavior can be analyzed in an unrivaled fashion. How the produced results can inform the design of driver assistance systems is presented next.

4 Using NDSs in Designing Driver Assistance Systems

In order to effectively improve safety and driver comfort, the design of driver assistance systems must follow a systems engineering process that centers on the drivers' needs. Systems engineering refers to the proactive, incremental design, and integration of system components to ensure that they work together. It opposes constructing each component separately and forcing them to fit together at the end (Chapanis 1996; Department of Defense Systems Management College 2001; Rivera and Karsh 2008). In following this approach, an optimal fit between the system's components can be achieved while controlling costs and using resources efficiently. In particular, the fit between the user and technology can be optimized by following a user-centered design approach in the systems engineering process, where the user (i.e., driver) is valued as the central system component throughout the design process.

Human Factors engineers are responsible for guiding the user-centered design process. They do this by performing the following tasks: (1) identifying user requirements early on in the design that the system must meet to be useful, (2) providing design input so that the system meets these user requirements, (3) selecting the appropriate method for testing candidate designs, and (4) selecting the appropriate analytical approach to evaluate whether the candidate designs meet the user requirements. Naturalistic driving research is a tool that can support the Human Factors engineer perform these duties during the design of driver assistance systems. The remainder of this chapter describes how NDSs can support driver assistance systems design.

4.1 Using NDSs to Identify User Requirements

The needs that a driver assistance system must meet to be effective are initially identified during its conceptualization. These needs are identified by analyzing crash databases, surveying users and domain experts, conducting interviews and focus groups with users and domain experts, performing task analyses, and performing work domain analyses (Barfield and Dingus 1998; Chapanis 1996; Sanders and McCormick 1993; Vicente 1999). Human factors engineers use the resultant information to develop user requirements to quantify how the new system must operate to meet drivers' needs (Chapanis 1996; Department of Defense Systems Management College 2001). The user requirements are then transformed into system requirements so that they can be easily followed by engineers working on the technological components.

Existing NDS databases offer a wealth of data that can be mined to assess what drivers need from a driver assistance system in order for it to improve their safety and comfort. For instance, a vehicle manufacturer may wish to know what types of crashes are most common so that appropriate countermeasures can be developed. Although this information can be acquired by analyzing crash databases, such as the National Automotive Sampling System General Estimate System (U.S. Department of Transportation 2009), epidemiological studies using such databases are unable to assess what the driver did leading up to the crash. This lack of information can lead to misclassification of events,

and may even mislead designers. As an example, it was once thought that following too closely was a common crash type. It was therefore believed that a driver assistance system that notified the driver when they were following too closely would reduce the occurrence of these types of crashes. However, there have been serious crashes recorded in NDSs that were erroneously coded in the police report as occurring because the driver was following too closely, but actually resulted from the driver being visually distracted. The drivers in these cases was actually following at a normal headway, but looked away from the forward roadway for an extended period of time. For this reason, the analysis of existing NDS datasets can shed insight on the magnitude of crash and near-crash events and what type of driver assistance system is best suited to reduce their occurrence. With respect to this example, a driver assistance system that notifies the driver of long glances away from the forward roadway may be a more appropriate countermeasure than a system that provides feedback to drivers regarding their headway. A forward collision warning system that notifies drivers of closing objects might also have served the drivers in these crashes.

With the magnitude of various crash types identified, a vehicle manufacturer might develop a driver assistance system that addresses a specific crash type. Existing NDS datasets can be queried to produce a subset that consists of safety-critical events that pertain to the crash type of interest. These events can then be investigated to identify what the driver did in the seconds leading up to the safety-critical event. For instance, an eye glance analysis could be performed to assess where the driver looked, or did not look, prior to the safety-critical event. This information could then inform the design of a new driver assistance system. For example, a specific type of lane change safety-critical event that has been observed consists of a driver swerving to avoid a suddenly decelerating lead vehicle (Fitch et al. 2009). These drivers are pressed to change lanes due to the urgency of the situation, and are less likely to use their turn signals or check their blind spots when executing a swerve avoidance maneuver. Such information might suggest that a lane change warning system should generate a non-visual alert, as drivers would not have time to perceive a visual alert during such events. Furthermore, the average amount of time that elapses from the driver initiating an avoidance maneuver to the safety-critical event may provide insight on how early the non-visual alert needs to be presented to be effectively processed by the driver. Fitch et al. (2009) found that lane change near-crashes resulted on average 1.8 s from the driver initiating a steering maneuver. This might suggest that a lane change warning system may best serve drivers if it provides an alert once a steering input is made, rather than waiting for the vehicle to cross the lane markings into the adjacent lane. This is because the alert may be too late to be effective if it is activated by a lane crossing.

4.2 Using NDSs to Guide the Design of Driver Assistance Systems

Human Factors engineers play a key role in the conceptual design of new driver assistance systems. They are responsible for reviewing current research and developing an understanding of the issues facing drivers when they interact with technology. They use this

information to develop design criteria that is then discussed with engineers responsible for the technological subsystem. A profound question that human factors engineers must address is whether a driver assistance technology will negatively impact driver performance. NDSs can help designers address this question because they have investigated the risk of driver interaction with technology in the real world on a subtask level.

NDSs have investigated the risk of performing various non-driving tasks, in particular interacting with mobile devices, in real-world driving conditions. NDSs have shown that “using a cell phone” in general significantly increases the risk of encountering an SCE. However, numerous NDSs that studied light and commercial vehicle drivers have also found that this increased risk is attributed to the complex subtasks pertaining to cell phone use, such as reaching for the device, dialing, and texting. In contrast, conversing on a hand-held or hands-free cell phone has not been found to increase risk, and has even been shown to significantly reduce risk. These results have been found when investigating the risk of mobile device use in low, moderate, and high driving task demands. Overall, this body of research suggests that it is not technology in the vehicle that impacts safety, but rather, how the driver interacts with the technology that implicates safety. A driver assistance system that eliminates complex subtasks and allows drivers to interact with the system while keeping their eyes on the road and their hands on the wheel may mitigate distracted driving based on the NDS findings. Voice controlled driver assistance systems stand to address this user requirement as they have been found to reduce the amount of time spent looking away from the forward roadway and the mental workload in performing a task (Owens et al. 2010). However, it is equally important that drivers do not look at the source of the audio information (i.e., speakers) for an extended period of time if the driver assistance system does not recognize their verbal command. As researchers continue to investigate driver distraction using NDS datasets, the results stand to provide meaningful insight for designers of driver assistance systems.

4.3 Using NDSs to Test Driver Assistance Systems

The proactive, incremental design of driver assistance systems requires iterative and frequent testing of design prototypes. Initial prototypes should be low in fidelity so that they are malleable and inexpensive to alter. These prototypes are tested to generate feedback on whether they meet the user requirements. Test results indicate where the system succeeded or failed to support drivers. Designers then modify the prototype to prevent the failures from reoccurring. Various alternative prototypes are typically generated during this process. The updated prototypes are then retested, restarting the design cycle. The best prototype is then selected and advanced in fidelity. By performing user testing early on in the design cycle, changes can be made before the system has undergone substantial development, drastically reducing the developmental costs and timeline. How NDSs can support prototype testing at various stages of the design process is presented below.

4.3.1 Testing Early-Stage Prototypes

A Collision Avoidance System's (CAS) effectiveness is highly dependent on how well it reliably detects unfolding conflicts and minimizes false and nuance alerts (Barfield and Dingus 1998). False alerts occur when an alert is activated because of a faulty sensor, while nuance alerts occur when the CAS functions as designed, but the situation does not constitute a true crash threat (Barfield and Dingus 1998). McLaughlin et al. (2008) developed a method for quickly and inexpensively bench-testing CAS algorithms using existing NDS datasets. The method consists of the following steps: (1) selecting safety-critical events in the NDS dataset that are relevant to the CAS, (2) overlaying the CAS algorithms on the events to determine at what point the driver would have received an alert. This is possible because the CAS inputs are typically available in the NDS dataset, (3) assess whether the algorithm correctly detected the safety-critical events, (4) investigate whether the CAS alert's timing was appropriate. This is done by applying human response performance distributions to the data (e.g., using a BRT and deceleration profile distribution to assess the percentage of the driving population that would have stopped in time to avoid a rear-end crash had they responded to the CAS alert), (5) update the algorithm sensitivity to meet the desired objectives if the percentage of the driving population computed to avoid the conflict is too low, and (6) retest the updated algorithms in an iterative fashion.

The algorithms can also be applied to nonevent data to determine the frequency of false and nuance alerts on a vehicle-mile traveled basis. CAS algorithm sensitivity, as well as appropriate filter criteria, can be developed through this approach. Overall, the method is a low-cost and useful tool for adjusting CAS algorithms prior to constructing any of the hardware components.

4.3.2 Testing Working Prototypes

Once a driver assistance system prototype nears a working design, it becomes worthwhile to perform a field test. The NDS approach can support field testing by allowing the direct observation of drivers' use of the system in real-world conditions. This field test approach has been successfully performed numerous times and has produced tremendous insight on driver interaction with advanced driver assistance systems (Fitch et al. [in press](#); General Motors Corporation 2005; Hanowski et al. 2008; LeBlanc et al. 2006). A description of how the NDS approach was used to evaluate a driver assistance system is provided below.

Camera/Video Imaging Systems (C/VISs) help drivers monitor the areas around their truck by displaying live video captured from cameras mounted on the truck's exterior on displays mounted inside the cabin. They are a low-cost countermeasure to improper lane changes, turning maneuvers, and backing crashes, which account for 6.3% of all truck crashes (Fitch et al. 2011). Recognizing that C/VISs were becoming readily commercially available, the U.S. DOT contracted the Virginia Tech Transportation Institute to investigate their design so that their benefits could be maximized. C/VIS concepts were developed and feedback on their expected utility was gained from focus groups (Wierwille et al. 2008). Early-stage



■ Fig. 21.6

C/VIS monitors and cameras installed on a truck (Fitch et al. [in press](#))

prototypes were developed and user tested on a controlled test track (Wierwille et al. [2008](#)). An all-around system was then developed by combining various C/VISs together and user testing it on the same test track (Wierwille et al. [in press](#)). The C/VIS program culminated with the development of a roadworthy C/VIS (► [Fig. 21.6](#)) that was field tested to investigate drivers' performance with the technology (Fitch et al. [in press](#)). A review of the user-centered system design of the C/VIS is presented in (Fitch et al. [in press](#)).

To field test the C/VIS, an NDS was devised to record driver performance for 1 month without a C/VIS, and then for 3 months with a C/VIS. Schneider National, Inc. allowed VTTI to instrument six of their trucks with the C/VISs. Commercial drivers from their Winchester, Virginia operation were then recruited to operate these vehicles for 4 months. The NDS methodology outlined earlier in this chapter was followed. Drivers completed questionnaires regarding their driving performance and C/VIS utility every 2 weeks. An NDS dataset containing 412,417 km of driving data, spanning 5,161 h, and comprising 3.35 Terabytes was produced. A total of 277 unique SCEs were identified, and 2,012 lane change maneuvers were sampled. The results provided useful insight on drivers' usage of C/VISs and are presented as examples in the next section.

4.4 Using NDSs to Evaluate Driver Assistance Systems

To evaluate a driver assistance system, aspects of drivers' performance and behavior that pertain to the objective of the driver assistance system should be measured. That way, the results can indicate whether the system is effective, or whether it negatively impacts performance.

In evaluating the C/VIS's effectiveness, a measure of drivers' rate of involvement in safety-critical events was used. This measure was developed in Blanco et al. (2009) and was the primary measure of whether a C/VIS improved drivers' performance. It was hypothesized that operating a commercial motor vehicle with a C/VIS would allow drivers to develop spatial awareness with fewer glances to the West Coast mirrors, thus becoming more likely to look at the forward roadway and detect unfolding conflicts. The SCE rate was computed as the number of safety-critical events encountered by a driver in a condition divided by the total number of hours spent driving above 3 mph in that condition. The mean SCE rate in the Baseline condition was then compared to the mean SCE rate in the Test condition using inferential statistics. For this study, a significant difference was not observed. The mean SCE rate was 5.2 SCEs per 100 h of driving in both the Baseline and Test conditions. Although an improvement in drivers' performance was not observed, it was reassuring that the C/VISs did not distract drivers and jeopardize their safety.

Another aspect of driving performance that was of interest was the clearance between the tractor-trailer and a vehicle trailing in the adjacent lane when a participant performed a lane change. This is because there was a concern that some drivers may abuse the enhanced visual information by cutting in closer to adjacent vehicles when vying for road position. Rear-facing radar units allowed the clearance to be precisely measured in the 2,012 sampled lane changes. It should be noted that the lane changes had to be made in front of an adjacent vehicle and when traveling above 35 mph to be included in the sample. It was found that drivers mean clearance to an adjacent vehicle when changing lanes did not significantly differ between the Baseline and Test conditions. To further investigate lane change performance, the clearance data was also binned so that the frequency of close lane changes (i.e., less than 6 m of clearance) could be investigated. Again, it was found that drivers did not execute significantly more "close" lane changes when driving with a C/VIS.

A primary concern when adding a visual display inside a motor vehicle is that drivers may decrease their likelihood of looking at the forward roadway and fail to perceive unfolding conflicts. Whether the C/VIS was a visual distraction was assessed by analyzing the eye glance data captured prior to a safety-critical event and a lane change maneuver. For this study, it was found that drivers did not become less likely to look forward in the 8 s preceding a safety-critical event when driving with a C/VIS. Drivers also did not become less likely to look forward in the 13 s surrounding a lane change maneuver when driving with a C/VIS. This finding, in addition to the findings on drivers' safety-critical event rate and lane change clearance, helped determine that the C/VIS did not negatively affect drivers' performance.

The eye glance analysis did reveal, however, that drivers were more likely to use the C/VIS when making right lane changes. This finding was reinforced by their subjective ratings, in which they indicated that the right blind spot was larger than the left blind side, and that the center and right C/VIS views were the most useful than the left C/VIS view. Future research will use the eye glance data to assess C/VIS usage patterns, such as whether drivers were more likely to look at the C/VIS display first, or the West Coast mirror first,

when preparing to change lanes. Computing the transition probabilities could help designers assess whether the monitors should be located closer to the West Coast mirrors (such as on the A-pillars), or closer to the drivers' fovea centralis when they are looking at the forward roadway.

On a final note, test criteria should be established to assess whether a driver assistance system has accomplished its goal. For example, had the number of lane changes in which the clearance was less than 6 m to an adjacent vehicle significantly increased when driving with a C/VIS, the design of the C/VIS, or how drivers are trained to use it, might need to be rethought. Had drivers reduced their probability of looking forward down to 0.66 when driving with a C/VIS (which is equivalent to 2 s of eyes-off-road time computed over a 6 s interval), the concern that the monitors impose a significant visual distraction would be substantiated. This could be argued because Klauer et al. (2006) found that eye glances totaling 2 s or more over a 6 s interval double drivers' risk of encountering a safety-critical event. Test criteria can therefore help designers identify when the system is production ready.

5 Conclusion

Naturalistic driving research facilitates the systematic execution of representative experimental design, bridging the gap between epidemiological and empirical research. Because NDSs capture drivers' in situ behavior and performance, the method and produced datasets are highly useful for the design, testing, and evaluation of driver assistance systems. This chapter described how NDS datasets can support the development of user requirements early on in the design of new driver assistance systems. It also outlined how the NDS method can be used to test candidate prototypes. Finally, it described how evaluation criteria unique to NDSs can be used to assess whether the candidate system improves driver performance, or whether it negatively impacts transportation safety. As the equipment needed to perform NDSs becomes cheaper, easier to install, and more readily available, naturalistic driving research will serve systems development into the future.

6 Case Study

The following case summarizes the approach described in this chapter. After performing a naturalistic driving study of commercial motor vehicle drivers, a serious crash was recorded in which a CMV driver executed an improper right lane change on a two-lane divided highway (Blanco et al. [in press](#)). Although the driver looked at her right West Coast mirror five times in the 15 s prior to steering right, she failed to see a light vehicle traveling in her front-right blind spot (the area to the right of the truck cab). The light vehicle was struck by the truck during the course of the lane change, causing the light vehicle to depart the road and crash. Fortunately, the driver recovered from the inflicted injury.

This event brought significant attention to the visibility issues facing CMV drivers. The data showed that despite drivers scanning the roadway, surrounding objects can be missed. It also showed that drivers can fail to use their fender-mounted mirrors, which would have showed the adjacent vehicle in this case had they been used.

Around this time, research was being performed at VTTI to determine the optimal design of Camera/Video Imaging Systems (C/VISs). C/VIS user requirements were being updated for a system that would allow drivers to perceive the areas around their truck and in inclement weather conditions (Wierwille et al. [in press](#)). Whether visibility provided by a C/VIS should overlap the visibility provided by the mirror system was in question. The lane change crash described above ultimately helped demonstrate to designers that it is worthwhile for C/VIS visibility to be redundant with the mirror system. An all-around system was then developed that consisted of C/VISs mounted on the left and right fenders, and a third wide-angle look-down C/VIS mounted on the top rear of the trailer (Wierwille et al. [in press](#)).

After user testing the C/VIS prototype on a controlled test track, there was a need to investigate how CMV drivers would use the system when driving on revenue-producing trips. A field test was thus performed using the NDS approach (Fitch et al. [2011](#)). DASs were installed in a fleet of trucks at Schneider National's Winchester, Virginia operation. CMV drivers were recruited to participate in the study. C/VISs were installed on the trucks and deactivated for the first month of participation so that drivers' performance without a C/VIS could be measured. The C/VISs were then enabled for 3 months, allowing the investigation of how they affect driver performance. Evaluation criteria specific to NDSs was then used to help assess drivers' performance. The study showed that drivers' rate of encountering safety-critical events did not change when driving with a C/VIS, indicating that the C/VIS did not disbenefit drivers. The study also showed that C/VISs assuage the workload pertaining to lane change maneuvers, particularly merge maneuvers. Drivers reported having improved spatial awareness when driving with a C/VIS. Drivers also did not reduce the time spent looking at the forward roadway, a major concern when placing new displays inside a vehicle. The study did provide insight on the importance of allowing drivers to reduce the display brightness at night and not block drivers' view of the fender-mounted mirrors (findings that surfaced by also investigating drivers' performance with a commercially available C/VIS). Overall, the study generated significant knowledge of "real-world" driver performance with a new driver assistance system.

The field test results are intended to help C/VIS manufacturers improve their product to better support CMV drivers. Because the data were collected over a year, there lies an opportunity to further investigate driver performance in specific conditions and when executing specific maneuvers. Because video of the drivers were continuously recorded, drivers' visual behavior with and without the C/VIS in these conditions can be examined in detail. The findings could provide additional feedback to designers on how C/VISs can better support drivers. The ability to mine the NDS dataset to answer future research questions that pertain to the design of driver assistance systems is truly a powerful feature.

References

- Barfield W, Dingus TA (1998) Human factors in intelligent transportation systems. Lawrence Erlbaum, Mahwah
- Blanco M, Bocanegra JL, Morgan JF, Fitch GM, Medina A, Olson RL, Hanowski RJ, Daily B, Zimmermann RP, Howarth HD, Di Domenico TE, Barr LC, Popkin SM, Green K (2009) Assessment of a drowsy driver warning system for heavy-vehicle drivers: final report. NHTSA Contract No. DTNH22-05-D-01019, Task Order #18. National Highway Traffic Administration, Washington, DC
- Blanco M, Hickman JS, Olson RL, Bocanegra JL, Hanowski RJ, Nakata A, Greening M, Madison P, Holbrook GT, Bowman D (in press) Investigating critical incidents, driver restart period, sleep quantity, and crash countermeasures in commercial operations using naturalistic data collection. Contract No. DTFH61-01-C-00049, Task Order # 23. Federal Motor Carrier Safety Administration, USDOT, Washington, DC
- Chapanis A (1996) Human factors in systems engineering. Wiley, New York
- Department of Defense Systems Management College (2001) Systems engineering fundamentals. Defense Acquisition University Press, Fort Belvoir
- Dingus TA (2003) Human factors applications in surface transportation. *Front Eng* 8:39–42
- Dingus TA (2008) Naturalistic driving: need, history and some early results. Retrieved 8 Nov 2010 from http://www.vtti.vt.edu/PDF/ndmas_ppt_PDFs/dingusVTTI.pdf. Accessed 15 June 2011
- Dingus T, Klauer S, Neale VL, Petersen A, Lee SE, Sudweeks J, Perez M, Hankey J, Ramsey D, Gupta S, Busher C, Doerzaph Z, Jermeland J, Knippling R (2006) The 100-car naturalistic driving study, Phase II – results of the 100-car field experiment. National Highway Safety Administration (NHTSA), Washington, DC
- Fitch GM, Lee SE, Klauer S, Hankey JM, Sudweeks J, Dingus TA (2009) Analysis of lane-change crashes and near-crashes. Technical Report No. DTNH22-00-C-07007, Task Order 23. National Highway Traffic Safety Administration, Washington, DC
- Fitch GM, Blanco M, Camden M, Olson R, McClafferty J, Morgan JF, Wharton AE, Howard H, Trimble T, Hanowski RJ (in press) Field demonstration of heavy vehicle camera/video imaging systems: final report. Contract No. DTNH22-05-D-01019, Task Order #23. National Highway Traffic Safety Administration, Washington, DC
- Fitch GM, Blanco M, Camden MC, Hanowski RJ (2011a) Field demonstration of a camera/video imaging system for heavy vehicles. In: Proceedings of the society of automotive engineers commercial vehicle engineering congress and exhibition, Chicago, IL
- Fitch GM, Schaudt WA, Wierwille WW, Blanco M, Hanowski RJ (2011b) Human factors and systems engineering of a camera/video imaging system. In: Proceedings of the 18th World congress on intelligent transportation systems, Washington, DC
- General Motors Corporation (2005) Automotive collision avoidance system field operational test (ACAS FOT) final program report. Technical Report No. DOT HS 809 886. General Motors Corporation and National Highway Traffic Safety Administration, Warren
- Hammond KR, Stewart TR (2001) The essential Brunswick: beginnings, explications, applications. Oxford University Press, Oxford/New York
- Hanowski RJ, Blanco M, Nakata A, Hickman JS, Schaudt WA, Fumero MC, Olson R, Jermeland J, Greening M, Holbrook GT, Knippling RR, Madison P (2008) The drowsy driver warning system field operational test: data collection methods final report No. DOT HS 810 035, Washington, DC
- Hickman JS, Hanowski RJ, Bocanegra J (2010) Distraction in commercial trucks and buses: assessing prevalence and risk in conjunction with crashes and near-crashes. Report No. FMCSA-RRR-10-049. Federal Motor Carrier Safety Administration, Washington, DC
- Hickman JS, Knippling RR, Olson RL, Fumero MC, Blanco M, Hanowski RJ (in press) Heavy vehicle-light vehicle interaction data collection and countermeasure research project, Phase 1 – preliminary analysis of data collected in the drowsy driver warning system field operational test: Task 5, preliminary analysis of drowsy driver warning system field operational test data. Contract No. DTNH22-00-C-07007, Task

- Order 21. Motor Carrier Safety Administration, Washington, DC
- Klauer SG, Dingus TA, Neale VL, Sudweeks JD, Ramsey DJ (2006) The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data (No. DOT-HS-810-594). NHTSA, Washington, DC
- Klauer S, Holmes L, Harwood L, Doerzaph Z (in press) Toward development of design guidelines for connected vehicles systems: evaluation of display location and application type on driving performance. National Highway Traffic Safety Administration, Washington, DC
- LeBlanc D, Sayer J, Winkler C, Ervin R, Bogard S, Devonshire J, Mefford M, Hagan M, Bareket Z, Goodsell R, Gordon T (2006) Road departure crash warning system field operational test: methodology and results (No. UMTRI-2006-9-2). The University of Michigan Transportation Research Institute, Ann Arbor
- McLaughlin SB, Hankey JM, Dingus TA (2008) A method for evaluating collision avoidance systems using naturalistic driving data. *Accid Anal Prev* 40(1):8–16
- Olson RL, Hanowski RJ, Hickman JS, Bocanegra J (2009) Driver distraction in commercial vehicle operations: final report. Contract DTMC75-07-D-00006, Task Order 3. Federal Motor Carrier Safety Administration, Washington, DC
- Owens JM, McLaughlin SB, Sudweeks J (2010) On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players. *SAE Int J Passenger Cars Mech Syst* 3(1):734–743
- Rivera AJ, Karsh B-T (2008) Human factors and systems engineering approach to patient safety for radiotherapy. *Int J Radiat Oncol Biol Phys* 71(Suppl 1):S174–S177
- Sanders MS, McCormick EJ (1993) Human factors in engineering design, 6th edn. McGraw-Hill, New York
- Treat et al (1979) Tri-level study of the causes of traffic crashes: final report. Volume I: causal factor tabulations and assessments: institute for research in public safety, Indiana University
- U.S. Department of Transportation (2009) National automotive sampling system general estimates system. Retrieved Nov 2010 from http://www.nhtsa.gov/people/nca/nass_ges.html
- Vicente K (1999) Cognitive work analysis: towards safe, productive, and healthy computer-based work, vol 1. Lawrence Erlbaum, Mahwah
- Wierwille WW (1981) Statistical techniques for instrument panel arrangement. In: *Proceedings of the NATA conference series III. Human Factors*, New York, pp 201–218
- Wierwille WW, Schaudt WA, Fitch GM, Hanowski RJ (2007) Development of a performance specification for indirect visibility systems on heavy trucks. Paper Number 2007-01-4231. *SAE Trans J Commer Vehicles* 2(116):264–275
- Wierwille WW, Schaudt WA, Spaulding JM, Gupta SK, Fitch GM, Wiegand DM, Hanowski RJ (2008) Development of a performance specification for indirect visibility systems in heavy vehicles final report supporting research. Report No. DOT HS 810 960. U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC
- Wierwille WW, Schaudt WA, Blanco M, Alden A, Hanowski RJ (in press) Enhanced camera/video imaging systems (e-c/viss) for heavy vehicles: final report. Contract No. DTNH22-05-D-01019, Task Order 6 (Submitted Sept 2008). U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC

22 Intelligent Speed Adaptation (ISA)

Jeremy J. Blum¹ · Azim Eskandarian² · Stephen A. Arhin³

¹Computer Science, Penn State University Harrisburg,
Middletown, PA, USA

²Center for Intelligent Systems Research, The George Washington
University, Washington, DC, USA

³Civil Engineering, Howard University, NW, Washington,
DC, USA

1	<i>Introduction</i>	583
2	<i>A Taxonomy of Intelligent Speed Adaptation Systems</i>	584
2.1	Speed Limit Calculation	585
2.2	System Output	585
3	<i>Benefits of Intelligent Speed Adaptation Systems</i>	586
3.1	Relationship Between Speed and Safety	586
4	<i>Speed Reduction and Benefits of ISA</i>	587
5	<i>Challenges Facing Intelligent Speed Adaptation Systems</i>	588
5.1	Acceptability and Effectiveness	588
5.2	Negative Safety Effects	589
6	<i>Research Case Study in the United States: Advanced Vehicle Speed Adaptation System</i>	590
6.1	Development of AVSAS	590
6.2	Results of Data Analysis and Consumer Acceptance	592
6.3	Conclusions	594
7	<i>Other Case Studies</i>	594
7.1	STARDUST Project	595
7.2	Swedish Test (Eslöv)	595
7.3	Managing Speeds of Traffic on European Roads (MASTER)	596
7.4	The Australian TAC SafeCar Project	597

7.5 INFATI Danish ISA 597

7.6 Swedish ISA Tests 597

7.7 The Netherlands ISA Experiment 598

7.8 ARENA Programme (Gothenburg, Sweden) 599

7.9 Flemish ISA 599

8 Conclusion 600

Abstract: Intelligent Speed Adaptation (ISA) systems are in-vehicle systems designed to improve driver compliance with safe speeds. These systems can provide information on safe speeds to driver, warn the driver when they are exceeding this limit, or control brakes or throttle to prevent speeding. Because of the link between excessive speeding and severe crashes, ISA systems have been called “the most powerful collision avoidance system currently available” (Carsten and Tate 2001). However, ISA systems do face challenges to their widespread deployment. Perhaps the most significant of these challenges is finding an appropriate balance between user acceptability and system effectiveness. The more effective that an ISA system is at reducing speeding, the less likely it is to be acceptable to drivers, particularly those who would benefit the most from ISA systems. In addition, some researchers have expressed concern about potential negative safety implications of ISA systems including driver unloading, driver distraction, negative behavioral adaptations, and negative interactions with other road users. This chapter presents an overview of ISA system configurations, potential benefits of ISA systems, and challenges faced by the systems. In addition, case studies, including large-scale field tests of ISA systems, are presented.

1 Introduction

Inappropriate speed is consistently cited as a significant contributing factor in serious vehicle crashes. Between 1995 and 2002, speeding was listed as a contributing factor in 40% of fatal crashes in Australia (Paine et al. 2007). Similarly, the National Highway Traffic Safety Administration estimates that speed is a contributing factor in approximately 31% of all fatal crashes in the United States in 2003 (NHTSA 2003–2006).

Analyses of driver behavior have identified a range of reasons for continued speeding, despite the dangers of speeding. Although reasons for speeding vary widely, most reasons fall into four categories (Paine 1996). A small number of speeders can be classified as risk takers, who routinely drive at excessive speeds. The remaining reasons for speeding account for the vast majority of speeders, with speeders falling into three roughly equal groups. Reluctant speeders would rather not speed, but do so because of external pressures including time pressure, a desire to travel at the prevailing speed of traffic, or due to intimidation by tailgating. Intentional speeders typically drive 10–15 km/h over the speed limit, often believing that this is a safe speed and the risk of a speeding ticket is low at these speeds. Inadvertent speeders do not realize that they are speeding, perhaps due to ignorance of the current speed limit or due to driving an unusually powerful vehicle or one with a smooth ride.

Enforcement and infrastructure-based approaches have had limited success in addressing the problem of speeding. Police enforcement and automated enforcement of speed limits do cause drivers to reduce their speed. However, these effects have often been found to be limited both to the specific area of enforcement action and to the specific times when the enforcement action is occurring (Teed et al. 1993). Likewise, roadway designs that include traffic calming measures also have the ability to reduce

instances of speeding, but typically only for a limited area nearby the location of the measures (Comte et al. 1997).

Unlike the enforcement and infrastructure-based approaches, Intelligent Speed Adaptation (ISA) has shown promise in addressing the problem of speeding. ISA systems are in-vehicle systems which warn and/or restrict drivers from driving at excessive speeds. ISA system designs vary widely in how they calculate appropriate speeds, how they warn the drivers, whether they can be disabled, and how and whether they take control actions to prevent excessive speeding.

This diversity in ISA systems shows promise in addressing excessive speeding, regardless of the underlying reasons for the speeding. In addition to promoting safer roadways by addressing excessive speeding, the widespread deployment of these systems has been postulated to provide environmental benefits and even improved system-wide travel times.

However, ISA systems must overcome significant challenges in order to realize these benefits. Perhaps, the most significant of these challenges is one of user acceptances. ISA studies have shown that drivers are far more likely to accept ISA systems that have a limited impact on their driving. On the other hand, ISA systems that provide mandatory enforcement of a maximum speed, while effective in curbing excessive speeding, tend to be rated as the least favorable by drivers. Moreover, drivers who would benefit the most from ISA systems tend to rate these systems the lowest.

In addition to user acceptance, there are concerns that the deployment of ISA systems may have a range of negative safety implications. Some studies have argued that the systems may lead to driver hypovigilance, in which drivers' attention to the driving task is lowered due to lowered demands for monitoring their speed. Other researchers have expressed concerns about some system designs leading to the opposite problem of driver overloading, in which the feedback from an ISA system may distract the driver. Moreover, there has been concern that the ISA systems may lead to negative behavior adaptations as drivers attempt to make up for lost time. Finally, there is also concern that if the ISA systems are deployed in only a subset of vehicles, these ISA systems may lead to negative interactions with other road users, for example, through increased tailgating of ISA-equipped vehicles.

The remainder of this chapter presents an overview of ISA systems. First, the taxonomy that is commonly used to characterize the range of ISA systems is presented. Then, the range of potential benefits of ISA systems is described. Subsequently, the concerns and challenges facing ISA systems are reviewed. The final sections present a number of case studies of ISA systems, followed by conclusions.

2 A Taxonomy of Intelligent Speed Adaptation Systems

Intelligent Speed Adaptation system can be classified based on the manner in which maximum speeds are calculated and the types of system output. System output includes warning modalities, the types of control actions, and the availability of system override.

As discussed in later sections, the configuration of an ISA has a significant impact on the system effectiveness and user acceptability.

2.1 Speed Limit Calculation

The activation speeds in ISA system can be fixed, variable, or dynamic (Carsten and Tate 2005). Early systems used fixed limits, in which one maximum speed is set. Whenever the vehicle exceeds this speed, the ISA system's actions are triggered, regardless of the speed limit where the vehicle is traveling. ISA systems with fixed limits have enjoyed widespread deployment in trucking fleets, which prevent trucks from being driven faster than 100 km/h (60–65 mph).

An ISA system with variable speed limits adjusts the system activation speed to match the speed limit on the current road. There are two principal methods that have been used by ISA systems to gain knowledge of the current speed limit. Vehicle-to-roadside wireless communication has been used, with roadside beacons transmitting the current speed limit to the onboard systems (e.g., Almqvist 1998). Other systems have used GPS receivers and digital maps with speed limit information in order to determine the current system speed limit (e.g., Jimenez et al. 2008). The safe speeds in the maps can be a function of more than just the posted speed limits, taking into account roadway geometry, line-of-sight data, and lowered speeds upstream of road features, such as curves.

ISA system with dynamic speed limits offers even more flexibility. In addition to taking into account the current speed limit, the system activation speed is also determined by other factors including weather conditions or congestion (Peltola and Kulmala 2000). In order to improve the user acceptance of ISA systems, research has also been conducted to make the system activation speed a function of driving style and behavior (Arhin et al. 2008a).

2.2 System Output

When a vehicle exceeds the system activation speed, the output of ISA systems can take a number of different forms. The systems may provide warnings to driver, take control actions, or record an incident of excessive speed. The user interfaces which provide warnings to driver include visual, auditory, and haptic mechanisms. In addition, ISA systems may be mandatory or voluntary, with the driver able to override or turn off the latter type of system.

ISA system approaches include advisory, warning, automatic control, and recording, or a combination of two or more of these approaches. Advisory ISA systems passively provide the driver with information on the current posted speed limit. Warning systems go further, and provide warnings to the driver when the driver is exceeding the system activation speed. Automatic control systems take automatic control of the throttle or brakes in order to regulate speed. For example, some ISA systems will use active braking

when a vehicle has to slow down due to a change in speed limit or a steep slope. In addition, ISA systems may take control of the throttle. This control can be in the form of a dead-throttle, where depressing the gas pedal has no effect when ISA system is activated. Alternatively, the stiffness in the gas pedal can increase when the system is activated to make it harder to depress. Recording ISA systems log instances of excessive speeding profile, as well as other safety-related driving information, like instances of hard braking. These logs can be reviewed by fleet operators; vehicle owners, for example the parents in the case of teenage drivers; or law enforcement authorities.

ISA system user interfaces provide alarms and notifications to drivers via visual displays, auditory warnings, and haptic displays. Visual displays have included dash-mounted LED displays, head-up displays that display information projected onto the front windshield, and enhanced speedometers.

Haptic displays present warnings to the driver via means that the driver feels. The driver must be able to quickly connect the haptic display with the intended meaning of the warning. Consequently, some ISA systems have used haptic displays that included stiffening of the accelerator pedal, in order to provide a clear indication that the driver should ease up on the throttle. Physical rumble strips are used to indicate that drivers should slow down. Other systems have adopted this mechanism to indicate excessive speed. These systems use virtual rumble strips, which produce vibrations in the car similar to those caused by driving over a physical rumble strip.

ISA system can also be classified based on whether the systems can be disabled. Mandatory ISA systems do not allow the driver to turn off or otherwise override the system. On the other hand, voluntary ISA systems allow the system to be turned off via a switch or overridden, for example, with kick-down mechanism. In this kick-down mechanism, if the driver presses hard on the accelerator the system action is overridden.

3 Benefits of Intelligent Speed Adaptation Systems

Research since the 1960s has reported a strong relationship between inappropriate speed and crashes. This research has shown that deviation from average mean speed increases the incidence of crashes. In addition to reducing speed-related crashes, ISA systems promise other benefits including environmental benefits and aggregate travel times.

3.1 Relationship Between Speed and Safety

As early as 1964, research into the relationship between speeding and crashes showed a U-shaped relationship between crash incidence and vehicle speed on freeways (Solomon 1964). Crash incidence was at a minimum when vehicles were traveling at the prevailing traffic speed of 65 mph. The crash incidence rate increased with larger deviations from this mean, both as vehicles traveled increasingly faster than the mean speed as well as when they traveled more slowly. Other researchers found a similar relationship between crash

incidence and speed on a variety of roadways with different prevailing speeds (Spitz 1984; Blackburn et al. 1989; Bowie and Waltz 1994). One study in the United Kingdom, for example, found that drivers traveling more than 1.8 standard deviations above or below the mean traffic speed had a significantly higher potential crash rates (Munden 1967).

The increased risk associated with low traveling speeds has been found to be partially attributable to additional driving maneuvers, such as stopping, merging on to a highway, or slowing down for a turn. For example, a 1970 study in Indiana used loop-detectors to monitor speeds of 216 vehicles involved in 114 crashes on state highways with speed limits between 40 and 65 mph (West and Dunn 1971). After removing crashes involving vehicles that were turning, the incidence of crashes involving slow-moving vehicles was significantly reduced.

On the other hand, there is a clear relationship between the risk of severe crashes and excessive speed above the mean travel speed. This relationship is predicted by physics-based models of crashes and borne out by the analysis of crash data, with the probability of severe injury or fatality in a crash increasing with speeds above the average travel speed. For example, analysis of crash data found that the probability of a fatality in a crash was proportional to the speed increase to the fourth power (Jokschi 1993). Other researchers found that the probability of a fatality in a collision increases exponentially with the deviation from the average speed (O'Day and Flora 1982).

4 Speed Reduction and Benefits of ISA

The primary benefit of ISA systems are increased road safety due to reduced instances of speeding. In addition, research has also pointed to reduced fuel costs and shorter aggregate travel times.

In addition to reducing the number of crashes, the deployment of ISA systems will likely result in fewer severe crashes.

The effectiveness of ISA in reducing crashes will be a function of the ISA system configuration (Carsten and Tate 2005). Carsten and Tate project that ISA systems will reduce total crashes between 10% and 36% depending on system configuration, and reduce fatal crashes by between 18% and 59%. Mandatory systems were projected to reduce crashes by about twice the rate of advisory systems, which simply display the speed limits, or voluntary systems that drivers can disable or override. In addition, systems with dynamic speed activation levels are projected to produce significantly fewer accidents than those with variable limits, which in turn are projected to reduce crash rates modestly over systems with fixed limits.

While there is widespread agreement that reductions in speeding can produce safer roadways, the environmental benefits of ISA may take time to be fully realized. Fuel consumption savings are projected to be highest on urban roadways and lowest on highways (Carsten and Tate 2005). One of the key drivers for fleet operator's adoption of ISA is reduced fuel consumption. However, in ISA trials of passenger vehicles, the fuel consumption has not been significantly reduced (Regan et al. 2006). However, if vehicles

are designed to optimize performance with ISA systems, both fuel consumption and emissions should be reduced (Paine et al. 2007).

ISA appears to have at most a modest effect on individual travel time, and may improve travel times in the aggregate. Estimates of increased travel times based on simulations have indicated an increase in travel time of 2.5% (Carsten and Tate 2005). However, field tests of ISA systems have found that there are modest increases in travel times, ranging from no difference in travel times (Regan et al. 2006) to approximately a 10% increase (Várhelyi et al. 1998). Some microscopic vehicle simulation experiments indicate that increased deployment of ISA systems will increase travel times on the whole (Piao et al. 2004). However, other researchers speculate that because ISA systems will reduce in collisions and the resulting traffic jams, aggregate travel times may be reduced through the widespread deployment of ISA systems (Paine et al. 2007).

5 Challenges Facing Intelligent Speed Adaptation Systems

While ISA systems promise great safety benefits, these systems face significant challenges in order to gain widespread acceptance (Blum and Eskandarian 2006). The most significant challenge is balancing user acceptance of the system with the system effectiveness. The more effective an ISA system is in reducing speeding, the least likely it is to be acceptable. In addition, researchers have also expressed concerns about driver hypovigilance, driver distraction, negative behavioral adaptations, and negative interactions with other users.

5.1 Acceptability and Effectiveness

ISA systems face significant hurdles in user acceptance. There are system configurations that can make ISA systems more acceptable to users, such as voluntary ISA or advisory ISA. However, research into these systems has shown that they are less effective in limiting instances of speeding.

Generally, research has shown that mandatory ISA systems are the most effective ISA systems. However, these systems are the least accepted, especially by drivers who could benefit the most. Drivers who have used these systems have complained about higher mental demands (Lahrman et al. 2001), increased stress, and frustration and vulnerability (Carsten and Fowkes 2000; Lahrman et al. 2001). After driving in an equipped car in an experiment, between 38% and 70% of the subjects who had negative feelings for the system indicated they would not install a mandatory ISA system for their vehicle (Várhelyi et al. 1998).

Advisory and warning systems, on the other hand, have been found to have higher levels of user acceptance. In one field test, for example, almost 30% of these drivers did not lower their speed at all with the advisory and warning ISA system (Lahrman et al. 2001).

This higher level of acceptance comes at a cost of lower effectiveness compared to mandatory ISA systems (Comte et al. 1997; Carsten and Comte 1997; Várhelyi et al. 1998; Päätaalo et al. 2001). In addition, some drivers find even advisory or warning systems unacceptable. In the field operational test mentioned earlier, close to 40% of the subjects described the system as annoying (Lahrmann et al. 2001).

Voluntary systems, like advisory and warning systems, have higher user acceptance levels than mandatory systems. Voluntary systems, which can be turned off by the driver, also are significantly less effective than mandatory systems, because they are often disabled at the times when the systems would restrict speeding (Biding and Lind 2002).

5.2 Negative Safety Effects

In addition to the challenge of balancing effectiveness and acceptability, there is concern that ISA systems may have negative safety effects. These potential impacts include driver hypovigilance, driver distraction, negative behavioral adaptations, and negative interactions with other road users.

Hypovigilance, or driver underloading, is the loss of situational awareness that arises when a driving task is automated. ISA systems automate a portion of the speed regulation task, and in one ISA field test, about one third of the drivers reported that ISA reduced the attention that they paid to driving (Loon and Duynstee 2001). In other studies, evidence of hypovigilance as a result of ISA has been mixed. Hypovigilance can also result if a system does not have dynamic speed limits that change with driving conditions, or if the ISA system is not active on all roadways. For example, in field tests in which the ISA system covered only some roadways, drivers of ISA systems paid less attention to speed regulation tasks on roadways outside the covered area (Saad et al. 2004; Hjälm Dahl and Várhelyi 2004).

In addition to concerns about hypovigilance, driver overloading or distraction has been reported as a potential problem in ISA systems. This concern has been expressed in surveys of drivers who have never used ISA systems (Saad and Hjälm Dahl 2004) and in one quarter of the drivers who used mandatory ISA systems in a field test (Päätaalo et al. 2001). The type of alarm or alert system may have a strong influence on the degree to which an ISA system distracts or overloads a driver. The use of verbal warnings in an ISA system, for example, could contribute to a delayed reaction time in drivers to other stimuli (Kojima et al. 2004).

Researchers have also expressed concern about mixed behavioral impacts on drivers. The impact of ISA systems on yielding and merging behavior has been mixed, with some fields reporting improved yielding behavior (Hjälm Dahl and Várhelyi 2004) and others reporting deteriorations in yielding (Persson et al. 1993). The impact of ISA on speed regulation behavior is also mixed. Some studies found that drivers of ISA systems drive faster in curves and through intersections, perhaps to make up for time lost due to ISA speed limiting at other times (Persson et al. 1993). Other studies found smoother deceleration (Várhelyi and Mäkinen 2001) and no compensatory speeding in intersections (Várhelyi and Mäkinen 1998).

Another area of concern is mixed interactions between drivers of vehicles with ISA systems and drivers of non-equipped vehicles. Drivers of mandatory ISA systems report feelings of insecurity due to tailgating by non-equipped vehicles (Lahrmann et al. 2001; Päätaalo et al. 2001). When ISA systems prevented drivers from overtaking a slow-moving vehicle, drivers in ISA-equipped vehicles were more likely to engage in tailgating (Carsten and Fowkes 2000).

6 Research Case Study in the United States: Advanced Vehicle Speed Adaptation System

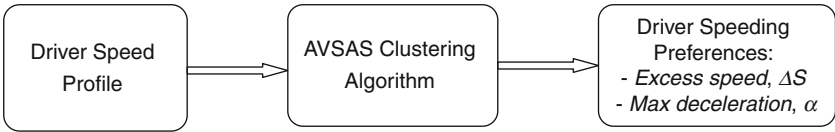
Between August 2006 and March 2007, the *Center for Intelligent Systems Research (CISR)* at The George Washington University (GWU) in the United States conducted a simulator experiment involving 21 drivers which comprised of four scenarios: baseline driving conditions (without any ISA), an advisory ISA, a mandatory ISA, and a new ISA called Advanced Vehicle Speed Adaptation System (AVSAS) (Arhin et al. 2008a, b). AVSAS was designed by CISR as a speed management system rather than speed limiting system based on individual driver speeding behaviors under normal driving conditions and different roadway scenarios.

6.1 Development of AVSAS

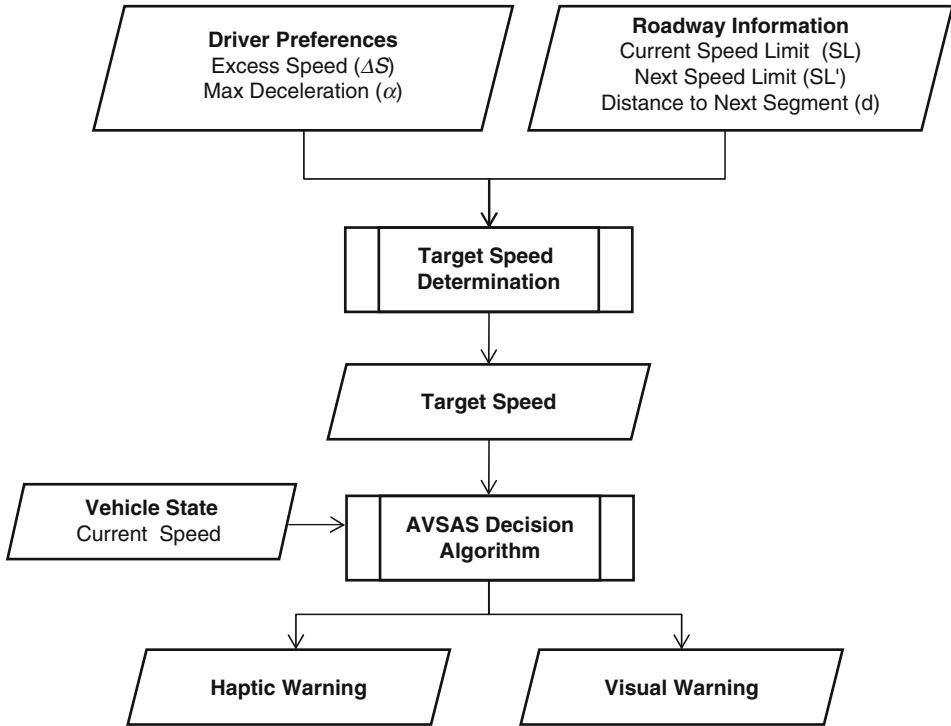
Drivers speeding behavior was analyzed and classified in the absence of any ISA system in order to identify different driving styles. Based on the analysis of the speed profiles of the drivers (excess speed above the posted speed limits, average speed, and maximum deceleration rates) without any ISA on different roadway classifications, three clusters of driver behaviors were developed. The clusters are conservative, normal, or aggressive.

► *Figure 22.1* presents the clustering algorithm developed for AVSAS.

Based on the clustering algorithm, the AVSAS speed regulation process was developed (see ► *Fig. 22.2*). The speed regulation process in AVSAS sets a preferred target speed based on the current roadway information for each driver. Thus, drivers within a particular cluster (conservative, normal, or aggressive) had a specific target speed for each roadway classification.



■ Fig. 22.1
Driver speeding behavior clustering



■ Fig. 22.2
AVSAS speed regulation process (Arhin et al. 2008b)

The Target Speed (TS) is defined as the sum of the roadway posted speed limit and the preferred excess speed above the posted speed limit. The excess speed above the posted speed limit depends on the classification of each driver.

Visual warning modes for AVSAS were also developed, samples of which are presented in ● Fig. 22.3. In this figure there are three different visual modes which provide color-coded-warning system based on the current speed of the driver. When the current speed is below the speed limit, no visual warning is given. If the driver's speed is between the posted speed limit and the preferred speed, an informative warning is provided. Finally, if the driver exceeds the preferred speed limit, a red zone of the excess speed is displayed which is mitigated by a haptic warning. The haptic warning, which was developed based on an algorithm for the AVSAS, is provided by increasing the magnitude of the force needed to press the gas pedal or to maintain its given position when the target speed is exceeded.

Another algorithm was also developed in AVSAS to enable drivers to have the option of overriding the haptic warning. This was developed with aim of providing drivers with the opportunity to drive above the preferred speed when absolutely necessary. The system can be overridden when the preferred or target speed is exceeded by pressing the gas pedal until the haptic system is disengaged.




Speed < Speed Limit	Speed Limit < Speed < Preferred Speed	Speed < Preferred Speed
No speed mitigation feedback	Speed excess marked by informative yellow zone, no active warning	Speed above Preferred Speed is shown by red zone and mitigated by haptic gas pedal
		

Fig. 22.3
Visual warning modes (Arhin et al. 2008b)

6.2 Results of Data Analysis and Consumer Acceptance

The AVSAS technology was assessed for its effectiveness and acceptance by conducting a series of statistical tests and a survey of the drivers who volunteered to take part in this experiment. To assess the effectiveness of the AVSAS technology, the following speed elements were analyzed:

- Maximum Speed (V_{max})
- Mean Speed (V_{ave})
- 85th Percentile Speed (V_{85})

The ANOVA test and the student’s t -test were respectively conducted to determine whether there are any statistically significant differences between the AVSAS technology with the baseline treatment to determine whether there are any statistically significant differences in these speed elements. The tests were conducted based on a 95% confidence interval.

The results of the ANOVA test showed that there is a statistically significant difference between the maximum, average, and 85th percentile speeds for the drivers under the baseline conditions and while using the AVSAS technology.

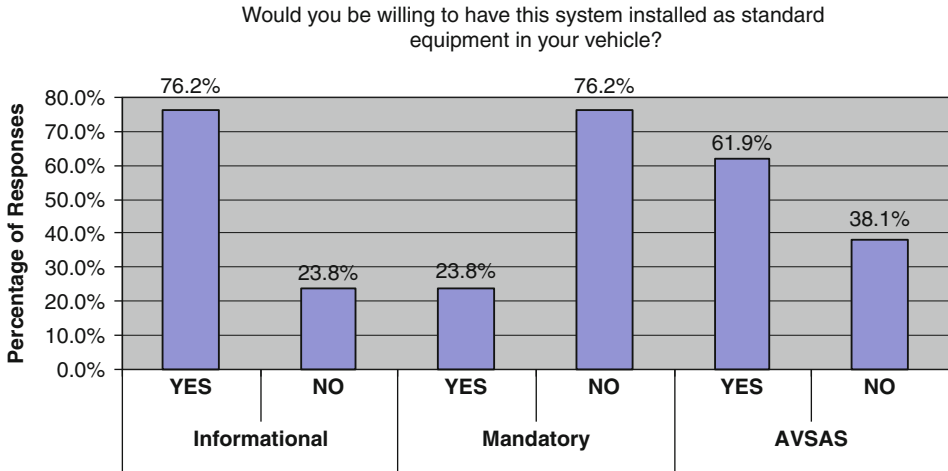
Of the six roadway classifications used in this experiment (local roads, collectors, arterials, freeways, residential and rural), the maximum speeds were reduced, on average by 67%, on four of them using the AVSAS technology. Using the student’s t -test, it was determined that the reductions were statistically significant for the four road classifications at 5% level of significance.

The mean speeds of all the drivers were reduced by approximately 3% on four of the six road classifications using AVSAS, compared with the baseline condition. These reductions, however, were not statistically significant at 5% level of significance ($p > 0.05$).

Of the six road classifications, the 85th percentile speeds were reduced the most on three using AVSAS by approximately 60%. These reductions in this speed measure were

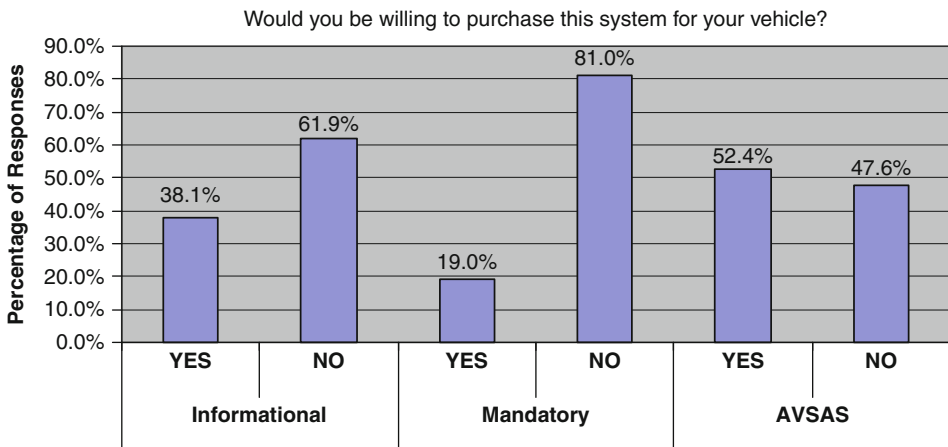
found to be statistically significant for only one of the three road classifications at 5% level of significance (urban: $p < 0.05$).

► [Figures 22.4–22.6](#) present the results of the survey conducted at the end of the experiment for all the ISA systems introduced, including AVSAS. Approximately 76% of the drivers reported that they would be willing to have the informational/warning ISA installed followed by AVSAS (~62%). The mandatory ISA was the least preferred.



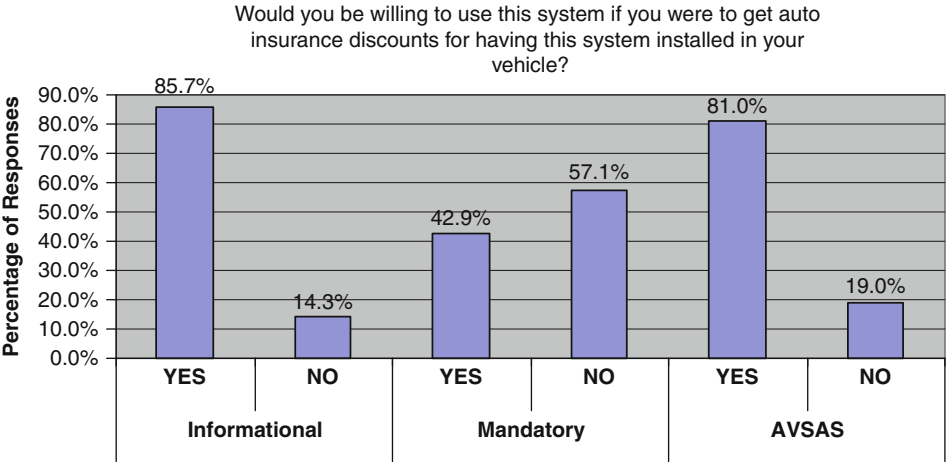
■ Fig. 22.4

Responses to “Would you be willing to have this system installed as standard equipment in your vehicle?” (Arhin et al. [2008a](#))



■ Fig. 22.5

Responses to “Would you be willing to purchase this system for your vehicle?” (Arhin et al. [2008a](#))



■ Fig. 22.6
Responses to “Would you be willing to use this system if you were to get auto insurance discounts for having this installed in your vehicle?” (Arhin et al. 2008a)

From Fig. 22.5, the drivers indicated that they would prefer to buy the AVSAS technology for their vehicle. When asked which system they would prefer if discounts are provided for insurance purposes, the majority of the drivers (85.7%) indicated that they would prefer to have the Informational/Warning ISA (see Fig. 22.6).

6.3 Conclusions

From the analysis, the AVSAS technology has the capability of mitigating the incidence of speeding, especially on freeways, thereby improving roadway safety. The results of the survey also show that the technology seems acceptable to drivers, compared with the mandatory ISA. Drivers indicated from the survey comments that, the ability to override the AVSAS system in peculiar roadway situations or conditions enabled them to choose the technology over the mandatory system. Maximum and average speeds on freeways were respectively reduced by 5% and 3% using AVSAS.

7 Other Case Studies

A vast number of field tests have been conducted on the effectiveness of ISA systems and corresponding driver behavior and acceptance. In each of the tests reviewed, varying types of ISA systems were tested and are described in this section. In particular, the following ISA simulator and/or field operating tests were reviewed:

- STARDUST Project (European Union)
- Swedish ISA Test (University of Lund)

- MASTER Project (European Union)
- The Australian TAC Project
- INFATI Danish ISA Test
- Swedish ISA Test (Large Scale City Tests)
- The Netherlands ISA Experiment
- ARENA Programme (Gothenburg, Sweden)
- Flemish ISA Test (Belgium)

7.1 STARDUST Project

As part of the European Union project, STARDUST assessed the extent to which Advanced Driver Assistance Systems and Automated Vehicle Guidance systems contribute to urban safety, a simulation study was conducted to assess the impact of ISA on traffic flow and speeding (Piao et al. 2004). A simulation model, called AIMSUN, was used for this study with the assumption that the drivers become familiar with the ISA systems. In addition to evaluating the speeding characteristics of the traffic stream, the study also evaluated the impact on lane changing behaviors, travel time, time headway, and time to collision. The simulation involved a gradual introduction of ISA-equipped vehicles in the traffic stream until a complete ISA-equipped traffic stream was achieved.

The baseline (with 0% ISA-equipped vehicles) average speed over the entire link was determined to be 61.4 km/h with a standard deviation of 4.7 km/h. The simulation results showed a gradual reduction in the average speed as well as the standard deviation as the percentage of ISA-equipped vehicles increased. With 20%, 80%, and 100% ISA penetration, the average speeds reduced to 60.8, 59.9, and 59.6 km/h respectively. On the other hand, the travel time increased with increasing introduction of ISA-equipped vehicles in the traffic stream.

7.2 Swedish Test (Eslöv)

The University of Lund conducted road tests within the city limits of Eslöv, Sweden and used vehicles equipped with a static, mandatory ISA system using a haptic throttle (Almqvist 1998). A haptic throttle is a mechanical device that could be manipulated to increase the accelerator resistance when the mandatory speed limit is attained. In the experiment, the ISA system was activated by roadside beacons located on ten roadways used to enter the city. Twenty-five subjects participated in the trial. The accelerator resistance was increased when driver entered an area with a speed limit of 50 km/h. Driver surveys were conducted after 2 months of using the system. Seventy-five percent of the subjects thought positively of the system, and conformity to speed limits increased during the test period.

7.3 Managing Speeds of Traffic on European Roads (MASTER)

The Managing Speed of Traffic on European Roads (MASTER) project was conducted between 1995 and 1998 and involved simulator experiments with a target speed limit of 80 km/h followed by road tests aimed at evaluating different ISA systems (Carsten and Comte 1997; Várhelyi et al. 1998; Várhelyi and Mäkinen 1998). In the simulator experiment, in which 60 subjects participated, it was found that the mandatory ISA were significantly more effective than other systems in reducing instances of speeding while advisory systems were more acceptable. The experiment tested both advisory and mandatory ISA systems and traditional speed control measures including in-car advice systems, transverse bars, in-car speed limiter, and a variable message sign. The results of the experiment showed some significant negative driver behavioral effects which included an increase in safety-critical car following in speed limited conditions, particularly in urban areas, delayed braking behavior as well as an increase in the percentage of red-light violations with the ISA system (Várhelyi et al. 1998). The researchers measured baseline measures of effectiveness and compared them with those of the ISA. In the baseline experiment, 33% of the subjects were involved with red-light-running violations. With the fixed mandatory system, 53% red-light-running violations were recorded, while for the advisory ISA system, 46% was reported. There were also more collisions with the mandatory system than in the baseline.

Driver acceptance of the ISA system was low. The majority of the subjects surveyed indicated that if the system costs approximately £50, they would prefer not have any of the ISA systems installed in their vehicle. As in other studies, the mandatory active system was viewed as being less favorable than the advisory system.

The MASTER program conducted a field test after the simulator tests in Sweden, Netherlands, and Spain with 20–24 subjects in each country (Comte et al. 1997; Várhelyi et al. 1998; Várhelyi and Mäkinen 1998, 2001). The ISA systems used were mandatory and variable systems in areas with posted speed limits of 30, 50, 60, and 70–120 km/h. The mandatory system consisted of an in-car speed limiter that consisted of an active gas pedal that resisted the applied force when drivers intended to exceed the speed limit. The resistance, which gradually increased as the vehicle speed approached the posted speed limit, helped remind drivers of the potential of exceeding the speed limit during the experiment. It was determined that the speed limiter reduced speeds significantly on roads with speed limits from 30 to 70 km/h. However, due to heavy traffic volumes on roadways with speed limits between 80 and 90 km/h, no statistically significant reduction in speeding was observed. Using the system, speed limit compliance improved while travel time increased up to about 9%.

Similar to the results from previous experiments, driver frustration was prominent with the mandatory system. Approximately 30% of subjects thought that these systems should be mandatory with 59% advocating that the system should be voluntary. Eleven percent of the participants were completely against the deployment of an ISA system in vehicles. General acceptance of the system was lowest among young drivers. Also, a majority of the subjects

thought the systems would be ideal for urban and built-up settings. Approximately 66–80% of the subjects believed that the mandatory and active systems would be appropriate when activated in certain road and weather conditions including poor visibility, slippery roads, in areas of high pedestrian volumes, among others.

The analysis of the data revealed that the level acceptance of the system by the subjects varied from country to country (Várhelyi and Mäkinen 1998). In Spain, a majority of the subjects responded negatively (70%) to potentially installing the system in their vehicle. However, in Sweden, about 62% of the subjects were willing to install the system in their own vehicles. The results indicated a split among the subjects in the Netherlands.

7.4 The Australian TAC SafeCar Project

Field tests of an ISA system were conducted by the Monash University, Australia as part of the Australian TAC SafeCar Project in which the system comprised of a combination of ISA, seat belt reminders, and following distance warning (FDW) systems (Regan et al. 2006). The system contained a two-level warning system. The first warning is given when the driver travels 2 km/h or more over the posted speed limit. The system issued an auditory warning and showed a visual warning display of the prevailing speed limit. There is then an increased resistance in the accelerator pedal if these initial warnings of speeding are ignored for two or more seconds. The FDW system warns the driver if he is following the vehicle immediately ahead too closely. In all, the system provided up to six different visual warning displays which are dependent on the time headway or when the following distance decreases. Typically, the system issues a final visual warning when the time headway is below 2 s, with an auditory warning following when time headways is less than 1.1 s.

7.5 INFATI Danish ISA

A road test of an advisory ISA system was conducted in Aalborg municipality, Denmark (Lahrmann et al. 2001; Jensen et al. 2005). The advisory system consisted of an LED display with speed limit, a warning light, and a voice that repeated warning of exceeding the speed limit every 6 s. This was conducted by Aalborg University in cooperation with a private organization. Analysis of the data indicated that the ISA system lowered speeds by 3–9 km/h on average. The drivers used in this road test showed adherence to the speed limits in urban areas where the posted speeds are lower than in rural areas where higher posted speed limits are common. About 39% of the drivers found the system annoying. Only 11% of the subjects did not lower their speed limits.

7.6 Swedish ISA Tests

The Swedish National Road Administration, Vägverket conducted road tests in four Swedish cities, Borlänge, Lidköping, Lund, and Umeå between 1999 and 2001, which

involved over 5,000 vehicles (Vägverket 2002; Hjalmdahl and Várhelyi 2004). The ISA technologies used were the advisory and voluntary systems. In Umeå, 3,642 vehicles were used for the test with a variable advisory ISA system that had a display warning system. Communication with roadside beacons that transmitted speed information to the vehicle was the primary technology that drove the variable speed limit feature. In the city of Lidköping, 110 vehicles were equipped with voluntary, variable ISA systems with active accelerators and GPS systems while an additional 110 vehicles had advisory, variable ISA systems with information display, enabled by GPS and digital maps. Three hundred and fifty vehicles were equipped with the ISA system and tested in the city of Borlänge. The ISA system tested were advisory systems that used GPS and digital maps using the same variable speed limit technology. Finally, 284 vehicles were tested in Lund with a voluntary active, variable ISA enabled by GPS receivers and digital maps.

From the results, it was determined that the systems were most acceptable especially on roadways with lower posted speed limits as well as in built-up areas where safety is of great concern. Some of such areas include residential areas, day cares, and schools which recorded between 80% and 95% acceptance rates. The voluntary ISA systems effectively reduced the mean and variance of the speeds of the drivers than the advisory and informative systems. The authors indicated that speeding violations decreased on the whole, especially in Lund due to the deployment of the active ISA system. There were variations of reduction in speeds depending on the type of road the test was conducted. Over a period of time, drivers were found to be disinterested in using the system, fewer wanted to keep the system, while a considerable number of drivers did not like the sound alerts of the system. Less than 70% of the drivers in Lidköping reported to be interested in keeping the informative system, while about 47% of the drivers said they would like to keep the voluntary active system. Also, in both Borlänge and Umeå, fewer than 60% of the subjects in each city said they would like to keep the informative system.

The analysis of violations showed mixed results in this study since most of the drivers turned off the voluntary system at the point when it would have been effective in reducing speed violations. With the ISA equipment, about 22% of the vehicles driven on 70 km/h posted speed limit roads were found to be over the speed limit, with 15% and 27% respectively for roads with 50 and 30 km/h speed limits.

7.7 The Netherlands ISA Experiment

In 1997, the Netherlands Ministry of Transport, Public Works, and Water Management conducted a road test of an ISA system. The test involved 120 subjects in the borough of Tilburg, Netherlands (Vanderschuren and Witziers 1998; Besseling 1999; Duynstee et al. 2001). The ISA system deployed for testing included mandatory ISA system with a number of road scenarios with different speed limits. The test revealed that average speeds of the drivers were reduced due to the ISA system. In addition, about 25% of the drivers violated fewer traffic rules, and the same percentage also claimed that they kept a larger distance from other road users because of the system. Analyses of the results also

showed that about 33% of the drivers reported an increase in their attention to their normal driving tasks, while an equal percentage of the drivers indicated a reduction. Some drivers reported that the system may cause them to be distracted while driving.

It was also determined that in areas of higher posted speed limits, more drivers found the ISA to be acceptable, although the deployed system was not the best (Duynstee et al. 2001). About 27% of the test subjects preferred driving without the ISA, while only 24% preferred driving with the system. Drivers also reported in that study that they were irritated and embarrassed due to tailgating by other drivers. On the whole, 62% of all the test drivers preferred driving without the ISA to driving with the ISA. About 52% of the drivers perceived that the ISA system would rather promote and improve safety for pedestrians and bicyclists (Duynstee et al. 2001).

7.8 ARENA Programme (Gothenburg, Sweden)

The ARENA Programme was road test conducted on a 35 km mainly rural route around Lake Aspen, in Gothenburg, Sweden (Almqvist and Towliat 1993). This system was a hybrid combination of ISA and Adaptive Cruise Control (ACC). The test vehicles would automatically drive the speed limit unless driver applied brakes. The speed limits on the tested roadways were 50–70 km/h. This combination of ISA and ACC had an adverse safety affect. Drivers felt pressured to drive too fast in villages and on sharp curves.

7.9 Flemish ISA

The Flemish Road Authority, D'leteren, ACUNIA conducted a road test in Belgium in 2002–2003 (Vlassenroot 2004). One hundred subjects participated in this test of a voluntary, variable ISA system. In this experiment, test drivers from both public and private sectors drove vehicles equipped with an Active Accelerator Pedal (AAP) system in the city of Ghent during 1 year. The test area included all speed limits possible and both urban and highway environments.

In the first phase of this study the effects of AAP on drivers' speeding behavior, perception of speed limits, attitudes about speeding, and acceptance and adaptation to driving with AAP were evaluated. Results show that drivers found the system useful for road safety because it allowed them to pay more attention to other road users, to find driving to be more relaxing, to look less at the speedometer, and to help to maintain the right speed (Vlassenroot and De Mol 2004). Although improvements in speeding behavior were noted in all testing areas, drivers preferred to use the system on highways or outside urban area than in urban zones and felt that driving in 30 km/h limited areas with the AAP activated was too slow.

System acceptance was also promising, however, the voluntary nature of participation in that study most likely skewed these results. There were high levels of voluntary use of the AAP on highways. Also, 15 drivers wanted to keep the AAP in their vehicle after the experiment. In a second phase of this study, the questionnaires about ISA were addressed

to the general public. Respondents agreed on the importance of speed management in traffic safety. Implementing ISA systems in all cars was perceived as a good solution to reduce speed violations. ISA showed a potentially high acceptance among the general public as a traffic safety-enhancing tool.

8 Conclusion

Because of the link between speeding and severe collisions, Intelligent Speed Adaptation (ISA) systems may be the most cost-efficient way to improve roadway safety. These systems are available in a range of different configurations, both in terms of the way that safe speeds are calculated and the way that the systems interact with the user. Safe speeds can be calculated based on fixed limit for the system, a limit that is based on the current speed limit or recommended safe speed for a roadway segment, or a limit that takes into account dynamic factors such as congestion and weather. The systems can be advisory systems that simply display the current speed, warning systems that warn when the system speed is exceeded, or mandatory systems that control the throttle to prevent speeding. In addition, many systems are voluntary in the sense that they allow the driver to disable the system.

While ISA systems have enjoyed widespread deployment in fleet vehicles, there are significant challenges in terms of acceptability and effectiveness in private vehicles. Mandatory ISA systems, the most effective of the configurations, have the lowest levels of driver acceptance. Warning, advisory, or voluntary systems, while slightly more acceptable, are significantly less effective. In addition, there are concerns that ISA systems could result in driver hypovigilance, driver overloading, negative behavioral adaptations, and negative interactions between drivers of ISA-equipped vehicles and drivers of non-equipped vehicles.

Acknowledgments

This work was supported in part by the US Department of Transportation under Cooperative Agreement No. DTFH61-05-H-00013.

References

- Almqvist S (1998) Speed adaptation: a field trial of driver acceptance, behavior, and safety. In: Fifth World Congress conference on intelligent transport systems (ITS), Seoul
- Almqvist S, Towliat M (1993) Roadside information linked to the vehicle for active safety: Aspen track. Swedish National Road Administration, Gothenburg
- Arhin SA, Eskandarian A, Blum J, Delaigue P, Soudbakhsh D (2008a) Effectiveness and acceptance of adaptive intelligent speed adaptation systems. *Transp Res Rec*, J Transp Res Board 2086:133–139
- Arhin S, Eskandarian A, Blum J, Delaigue P (2008b) Development and evaluation of an advanced intelligent speed adaptation system. *Proc*

- Inst Mech Eng D J Automobile Eng 222(D9): 1603–1614
- Besseling JFM (1999) Intelligent speed adaptation: the Dutch experiment. Urban transportation systems, Lund
- Biding T, Lind G (2002) Intelligent speed adaptation (ISA): results of large-scale trials in Borlänge, Lidköping, Lund and Umeå during the period 1999–2002. Vägverket, Borlänge
- Blackburn RR, Moran R, Glauz WD (1989) Update of enforcement technology and speed enforcement devices. Report No. DOT-HS-807 584. National Highway Traffic Safety Administration, Washington, DC
- Blum J, Eskandarian A (2006) Managing effectiveness and acceptability in intelligent speed adaptation systems. In: IEEE intelligent transportation systems conference, Ontario, 17–20 Sept 2006, pp 319–324
- Bowie NN Jr, Waltz M (1994) Data analysis of the speed-related crash issues. Auto Traffic Saf 1(2):31–38
- Carsten O, Comte S (1997) UK work on automatic speed control. In: Proceedings of the international cooperation on theories and concepts in traffic safety (ICTCT '97) conference, Lund, 5–7 Nov 1997
- Carsten O, Fowkes M (2000) External vehicle speed control: final report. The University of Leeds and Motor Industry Research Association, Leeds
- Carsten O, Tate F (2001) Intelligent speed adaptation: the best collision avoidance system? In: Proceedings of 17th conference on the enhanced safety of vehicles, Amsterdam, 4–7 June 2001, pp 1–5
- Carsten O, Tate F (2005) Intelligent speed adaptation: accident savings and cost-benefit analysis. Accid Anal Prev 37:407–416
- Comte SL, Várhelyi A, Santos J (1997) The effects of ATT and non-ATT systems and treatments on driver speed behavior. Managing speeds of traffic on European roads, Working paper R 3.1.1 of the MASTER project
- Duynstee L, Katteler H et al (2001) Intelligent speed adaptation: selected results of the Dutch practical trial. In: The 8th World Congress on intelligent transport systems, Sydney, 30 Sept–4 Oct 2001
- Hjälmdahl M (2004) In-vehicle speed adaptation: on the effectiveness of a voluntary system, Bulletin 223. Lund University, Lund
- Hjälmdahl M, Várhelyi A (2004) Speed regulation by in-car active accelerator pedal: effects on driver behavior. Transp Res F 7(2):77–94
- Jensen C, Lahrmann H et al (2005) The INFATI data. TIMECENTER technical report, TR-79, Copenhagen
- Jimenez F, Aparicio F, Paez J (2008) Evaluation of in-vehicle dynamic speed assistance in Spain: algorithm and driver behavior. IET Intell Transp Syst 2(2):132–142
- Joks HC (1993) Velocity change and fatality risk in a crash: a rule of thumb. Accid Anal Prev 25(1): 103–104
- Kojima S, Uchiyama Y et al (2004) Evaluating the safety of verbal interface use while driving. R&D Rev Toyota CRDL 39(1):23–38
- Lahrmann H, Madsen JR et al (2001) Intelligent speed adaptation: development of a GPS-based ISA system and field trial of the system with 24 test drivers. In: Proceedings of the 8th World Congress on intelligent transport systems (ITS World Congress), Sydney
- Loon AV, Duynstee L (2001) Intelligent speed adaptation (ISA): a successful test in the Netherlands. In: Proceedings of the Canadian Multidisciplinary Road Safety Conference XII, London, Ontario, June 2001
- Munden JM (1967) The relation between a driver's speed and his accident rate. Report LR 88, Transport and Road Research Laboratory, Crowthorne
- NHTSA (National Highway Transportation Safety Administration) (2003–2006) Analysis of speed-related fatal motor vehicle traffic crashes. NHTSA, Washington, DC
- O'Day J, Flora J (1982) Alternative measures of restraint system effectiveness: interaction with crash severity factors, SAE technical paper, 820798. Society of Automotive Engineers, Warrendale
- Päätalo M, Peltola H et al (2001) Intelligent speed adaptation: effects on driver behavior. European Working Group on Speed Control, Aalborg
- Paine M (1996) Speed control devices for cars. NSW roads and traffic authority report, May 1996
- Paine M, Paine D, Griffiths M, Germanos G (2007) In-vehicle intelligent speed advisory systems. In: Proceedings of the 20th international conference on the enhanced safety of vehicles, Lyon, 18–21 June 2007
- Peltola H, Kulmala R (2000) Weather related intelligent speed adaptation: experience from a simulator. Technical Research Centre, Finland

- Persson H, Towliat M et al (1993) Speed limiter in the car: a field study on speeds, behavior, conflicts, and driver comments when driving in a built-up area. Department of Traffic Planning and Engineering, University of Lund, Sweden
- Piao J, McDonald M et al (2004) An assessment of ISA impacts through micro-simulation. In: The 7th international IEEE conference on intelligent transportation systems, Washington, DC
- Regan M, Triggs T, Young K, Tomasevic N, Mitsopoulos E, Stephan K, Tingvall C (2006) Onroad evaluation of ISA, following distance warning and seat belt reminder systems: final results of the TAC Safecar Project. Monash University Accident Research Centre, Clayton
- Saad F, Hjalmdahl M et al (2004) Information Society Technologies (IST) programme: literature review of behavioural effects. European Union's fifth framework programme for research and technological development, Luxembourg
- Solomon D (1964) Accidents on main rural highways related to speed, driver and vehicle. Federal Highway Administration, Washington, DC
- Spitz S (1984) Speed versus speed limits in California cities. *ITE J* 54(4):42–45
- Teed N, Lund AK, Knoblauch R (1993) The duration of speed reductions attributable to radar detectors. *Accid Anal Prev* 25:131–137
- Vägverket L (2002) Lund: results of the ISA study. Swedish National Road Administration, Borlange
- Vanderschuren M, Witziers K (1998) Intelligent speed adaptation: the Dutch experiment in an urban area. In: 5th World Congress on intelligent transport systems, Seoul, 12–16 Oct 1998
- Várhelyi A, Mäkinen T (1998) Evaluation of in-car speed limiters: field study. Managing speeds of traffic on European roads. MASTER Working Paper, 3.2.2, Brussels
- Várhelyi A, Mäkinen T (2001) The effects of in-car speed limiters: field study. *Transp Res C Emerg Technol* 9:191–211
- Várhelyi A, Comte S et al (1998) Evaluation of in-car speed limiters: final report. MASTER Project, UK
- Vlassenroot S (2004) Intelligent speed adaptation (ISA) in Ghent, Belgium: the first European trial with politicians, academics and car-constructors as role models in ISA-driving. In: Brebbia CA, Wadhwa LC (eds) Tenth international conference on urban transport and the environment in the 21st century, urban transport, Dresden. Wessex Institute of Technology, Southampton
- Vlassenroot S, De Mol J (2004) Trial on intelligent speed adaptation in Ghent, Belgium: the results on acceptance and driving-behavior of the test. In: Proceedings of 11th World Congress and exhibition on intelligent transport systems and services, Budapest, 24–26 May 2004. ITS Europe, Brussels
- West LB Jr, Dunn JW (1971) Accidents, speed deviation and speed limits. *Traffic Eng* 41(10):52–55

Safety and Comfort Systems

Werner Huber and Klaus Kompass

23 Safety and Comfort Systems: Introduction and Overview

Klaus Kompass · Werner Huber · Thomas Helmer
BMW Group, Munich, Germany

1	<i>Safety and Comfort</i>	606
2	<i>The Driver–Vehicle–Environment Control Loop</i>	607
3	<i>Human–Machine Interaction</i>	609
4	<i>Development Principles</i>	610

Abstract: In recent years, research and development activities of automobile manufacturers have placed an increasing focus on offering intelligent assistance systems in the vehicle. By delivering targeted information and warnings, by delegation of tasks, or by intervention, these functions aim to improve active safety, particularly in complex situations, and/or to enhance the driver's sense of comfort. The system of driver, vehicle, and environment can be thought of as a control loop with feedback, in which the role of human drivers is decisive in determining the safety potential of this control loop. The safety and comfort characteristics arise for the driver from his interactions with the vehicle and the environment (Bernotat 1970). The human-machine interaction serves as the interface between the driver and the vehicle and also between the driver and the environment. Thus, the design of the human-machine interface is a key determinant of the effectiveness and acceptance of driver assistance systems. Due to the increasing amount of systems in a car, the functional integration of different assistance systems and a higher degree of automation of the functions are expected in future.

1 Safety and Comfort

In recent years, research and development activities of automobile manufacturers have placed an increasing focus on offering intelligent assistance systems in the vehicle. By delivering targeted information and warnings, by delegation of tasks, or by intervention, these functions aim to improve active safety, particularly in complex situations, and/or to enhance the driver's sense of comfort.

In this context, the term “function” refers to realization of a defined driver assistance goal, independent of the technical solution implemented to achieve it (Ebner et al. 2009). Thus, typical examples of functions are “collision avoidance” or “lane keeping.” A particular system (e.g., Emergency Braking, Heading Control) represents the concrete technical implementation of a function and thus contributes to achieving the defined goal; a system is not simply a collection of technical components, but generally involves strategies and algorithms designed to implement the required function.

Safety: The motivation for concerted efforts to improve vehicle safety is clear: Every year, an estimated 1.3 million persons (WHO 2009) are killed in traffic accidents. In addition to the loss of human life, traffic deaths and injuries as well as damage to property also represent a significant economic loss; in some countries, the costs approach 2% of the gross national product (WHO 2004). At the same time, the desire for individual mobility in society and the resulting increases in traffic demand (total kilometers driven) pose additional challenges to improved traffic safety.

Comfort represents an improvement in the driving experience that is perceived and valued by customers. Comfort is generally associated with concepts such as convenience, satisfaction, enjoyment, or luxury (Bubb 2003). Aspects such as esthetics and pleasure also play a role in assessment of comfort. According to Herzberg (1958), it is useful to describe comfort as the absence of discomfort: Discomfort can be quantified on the basis of physiological and biomechanical factors, whereas comfort is associated with “pleasure,”

a characteristic that is subject to substantial individual variability and is more difficult to measure or quantify. In any case, as pointed out by Helander and Zhang (1997), comfort and discomfort are not mutually exclusive, but are best described as the extreme points along a continuous scale.

Driver assistance systems are intended to influence both comfort and safety positively. Indeed, comfort and safety also influence each other, as illustrated in the following example: Air conditioning is generally regarded as a comfort system. However, regulating the air temperature in a vehicle not only enhances individual perception of a pleasant climate, climate control also has an influence on human alertness and performance, so that safety enhancement results as a secondary effect. If the driver experiences a temperature that he perceives as pleasant or comfortable, his performance level will also be closer to his optimal capabilities, since human performance levels strongly deteriorate outside of a certain temperature range (Wenzel 1993). Thus, in this example, there is no conflict between comfort and safety objectives. Performance deterioration can affect driving and thus negatively influence safety. At the individual level, the lack of a clear distinction between safety and comfort is also reflected in customer perceptions (Belz et al. 2004; Fuhrmann 2006). The difference is that whereas comfort is essentially an individually perceived phenomenon, improved safety represents a collective challenge that the society as a whole needs to address.

2 The Driver–Vehicle–Environment Control Loop

The system of driver, vehicle, and environment can be thought of as a control loop with feedback, in which the role of human drivers is decisive in determining the safety potential of this control loop. Driver assistance systems are designed to support the driver in his driving task throughout a spectrum of driving situations and thus contribute in varying degrees to the safety and comfort of the vehicle in each situation.

The safety and comfort characteristics arise for the driver from his interactions with the vehicle and the environment (Bernotat 1970). The dynamical nature of these interactions can be illustrated by a control theory model (Donges 1978; Bubb 1993; Ehmanns et al. 2000). Here, the human driver represents a complex controller, selecting the route, target variables (such as car following gaps), and controller actions. The driver responds to inputs from the environment and to feedback from the vehicle. This basic controller scheme can be extended to include the element of driver assistance. Of course, driver assistance systems are part of the vehicle, but they differ from “standard” vehicle controls by interacting with the driver, with the standard vehicle controls, and with the environment (Naab 2000; Bubb 2001; Knapp et al. 2009). Driver assistance systems compile information on the vehicle, the environment, and the driver; assess and interpret this information using internal system models; and calculate a target behavior or response. If the current state deviates from the target, a driver assistance system will calculate the appropriate action or feedback to the driver. System control responses range from providing the driver with information or warnings to carrying out automatic

interventions in vehicle dynamics; the intensity of system control response depends on the design characteristics of the particular driver assistance system, the reliability of algorithms for interpreting and classifying the current driving state, and the assumed criticality of the situation. Safety and comfort arise as consequences of the well-coordinated interaction of all elements of the control loop.

Support for the primary driving task can be provided at any level of the following three-level model hierarchy (Bernotat 1970):

- The high-level navigation task arises from the need to achieve the desired objective of the trip and reach the required destination; it comprises route planning and trip time estimation, with possible adaptation of the route to traffic conditions.
- At the intermediate level, the driving control task requires derivation of target variables, such as lane choice and desired speed, taking into account boundary conditions and external influences such as the dynamics of traffic flow. Driving maneuvers are carried out in accordance with control requirements in order to fulfill the navigation task.
- The lowest level of the model represents the process of stabilization; it includes all tasks that keep the vehicle “on course” (e.g., steering and braking).

Driver assistance systems can be designed to provide support at any of these three levels of the primary driving task. The driving state can be continually monitored in order to generate corrections on any or all of these levels if required; detailed applications, variations, and refinements of this model can be found in the literature (Rasmussen 1983; Ehmanns et al. 2000; Bubb 1993; Kompass and Reichart 2006).

Since the driver’s role in this control loop is decisive, it is instructive to consider the characteristics of driver behavior in detail. A classical hierarchical behavior model for targeted actions has been described by Rasmussen (1983). This model distinguishes three categories of “cognitive demands on humans in work processes”: knowledge-based, rule-based, and skill-based behavior.

If a person is confronted with complex tasks requiring untrained actions or reactions, the cognitive demands result in “knowledge-based” behavior. In this behavioral mode, possible actions are first mentally reviewed before the strategy that appears to provide the best solution is implemented.

People will generally carry out “rule-based” behavior in situations that they have repeatedly experienced, drawing on an inventory of learned rules or behavior patterns. These readily available rules and patterns allow a faster response to the situation.

“Skill-based” behavior arises if situation demands have been trained in a learning process and stimulus-response mechanisms are characterized by reflexive actions. Responses and performance are fastest at this level due to the routine and essentially autonomous execution of processes and actions. People carrying out skill-based behavior normally have the greatest capacity for processing secondary tasks while driving. By taking the driving task requirements and the behavioral level into account, one can compare the time required by the driver with the time available to him for particular situations and maneuvers. This comparison facilitates estimation of the driver’s needs and potential benefits of an assistance system (Reichart 2001).

3 Human–Machine Interaction

Driver assistance has the potential to solve complex safety problems involving the interplay of human capabilities, technology, and the environment and in addition to increase enjoyment of driving by improving safety and comfort. The design of the human–machine interface is a key determinant of the effectiveness and acceptance of driver assistance systems. Interpreted in terms of the control loop model, the human–machine interaction serves as the interface between the driver and the vehicle and also between the driver and the environment.

We can distinguish two basic approaches to the driver's role and his integration into an assistance system: improved performance and reliability (Kompass and Huber 2006).

- Assistance approach: The driver remains as the central element in the control loop, whereby driver assistance is designed to increase his performance and reliability.
- Automation approach: The driver is considered to be the most error-prone element with respect to accident risk and is therefore temporarily bypassed in the control loop by partly or completely automating the driving task.

A user-adapted, comprehensive support concept is essential for the design of driver assistance systems with the aim of enhancing human performance. The development and effective optimization of an entire driver assistance system thus requires an interdisciplinary team of engineers, scientists, and psychologists.

In a general context, potential countermeasures to prevent accidents and enhance traffic safety need to address all elements of the control loop as well as global or external factors such as road infrastructure, regulations and enforcement, driver education, and vehicle construction techniques. Since the spotlight here is on intelligent vehicles, the following discussion focuses on vehicle-centered safety approaches.

Key design rules for driver assistance systems can be summarized as follows (Naab and Reichart 1998):

- Driver assistance systems should act as a virtual copilot and decrease the burden on the driver.
- Driver assistance systems should enhance the driver's competence.
- Driver assistance should maintain driver control by supportive rather than intrusive system design. The driver should not feel involuntarily controlled by the driver assistance system.
- Driver assistance systems at the maneuvering level must be designed so that the driver can override the system at any time.
- The operation of driver assistance systems should be easily mastered and as intuitive as possible.
- The driver should be able to switch a driver assistance system off or on.
- Driver assistance systems exhibit transparent system behavior and conform to expected system properties.
- The effort required to operate and monitor a driver assistance system should not exceed the intended decrease in the driver's burden.

- Driver assistance should aim for an intermediate level of driver activation, i.e., avoiding both overburdening and monotony.
- The introduction of multifunctional driver assistance must take the principle of smooth coordination of functions into account; perceived inconsistencies and discontinuities should be avoided or at least rendered controllable, and the intensity of assistance provided by different functions should be compatible, particularly when transitions between functions can occur.

4 Development Principles

In the development process for vehicle active safety systems, it is useful to pursue a user and problem oriented top-down approach focused on the fulfilling the desired safety requirements. The requirements on active safety systems derive from societal goals of improved traffic and vehicle safety as well as from legal and regulatory mandates and other constraints. At the same time, in the area of individual mobility, the value of comfort to customers is also an important consideration.

Going beyond minimum legal requirements for vehicle model approval, the automobile industry aims to establish and standardize guidelines for development, design, and assessment of driver assistance systems. For example, the RESPONSE Code of Practice (Knapp et al. 2009) for design and evaluation of driver assistance is used in Europe; its use is recommended by the European Automobile Manufacturers' Association (www.ACEA.be). In North America, the Alliance of American Manufacturers has issued a document in 2006 with guidelines for design of driver-vehicle communication; automobile manufacturers can commit themselves voluntarily to compliance with these guidelines (Alliance of Automobile Manufactures 2006). The automobile industry has recognized its responsibility for designing novel systems in order to provide maximum benefit to drivers and other affected persons.

As implemented, various driver assistance systems often address differing safety goals or place a differing degree of emphasis on particular objectives. Technical progress will enable realization of an increasing variety of functions. For customer-oriented implementation of these goals, functional and HMI integration is absolutely essential. In order to achieve the penetration rates needed for desired effects – particularly those involving improved safety – the customer's motivation for selecting vehicle options cannot be ignored. Focusing on the customer's needs includes taking into account regional differences: The nature of traffic problems as well as customer preferences can differ strongly from one region to another.

In the future, functional integration of different assistance systems and a higher degree of automation of the functions are expected.

In the following chapters, selected systems and functions will be presented in detail:

- [Chapter 9](#), “Adaptive and Cooperative Cruise Control”
- [Chapter 25](#), “Forward Collision Warning and Avoidance”

- Chapter 26, “Lane Departure and Lane Keeping”
- Chapter 27, “Integral Safety”
- Chapter 28, “Lane Change Assistance”
- Chapter 29, “Steering and Evasion Assist”
- Chapter 30, “Proactive Pedestrian Protection”
- Chapter 31, “Parking Assist”
- Chapter 32, “Post-crash Support Systems”
- Chapter 33, “Map Data for ADAS”

References

- Alliance of Automobile Manufacturers (2006) Principles, criteria and verification procedures for driver interactions in advanced vehicle information and communication systems. <http://www.autoalliance.org/files/DriverFocus.pdf>, Accessed 7 July 2010
- Belz J, Höver N, Mühlenberg M, Nitsche B, Seubert T (2004) Fahrerassistenz im Spannungsfeld zwischen Komfort- und Sicherheitsanforderungen. Integrierte Sicherheit und Fahrerassistenzsysteme, VDI-Berichte 1864, Düsseldorf, pp 441–468
- Bernotat R (1970) Anthropotechnik in der Fahrzeugführung. Ergonomics 13:353–377
- Bubb H (2001) Haptik im Kraftfahrzeug. In: Jürgensohn T, Timpe KP (eds) Kraftfahrzeugführung. Springer, Berlin/Heidelberg/New York
- Bubb H (2003) Fahrerassistenz – primär ein Beitrag zu Komfort oder für die Sicherheit? Der Fahrer im 21. Jahrhundert. VDI-Berichte 1768, Düsseldorf
- Bubb H (1993) Informationswandel durch das system. In: Schmidtke H (ed) Ergonomie. Carl Hanser Verlag, München
- Donges E (1978) Ein regelungstechnisches Zweiebenen-Modell des menschlichen Lenkverhaltens im Kraftfahrzeug. Z Verkehrssicherheit 24:98–112
- Ebner A, Helmer T, Huber W (2009) Bewertung von Aktiver Sicherheit - Definitionen, Referenzsituationen und Messkriterien 1. Automobiltechnisches Kolloquium; München, 16 und 17. April 2009, Technische Universität München-Garching, VDI Wissensforum GmbH, 2009
- Ehmanns D, Gelau C, Nicklisch F, Wallentowitz H (2000) Zukünftige Entwicklung von Fahrerassistenzsystemen und Methoden zu deren Bewertung. 9. Aachener Kolloquium Fahrzeug- und Motorentechnik
- Fuhrmann KH (2006) Fahrerassistenzsysteme: Komfort und Sicherheit Integrierte Sicherheit und Fahrerassistenzsysteme. VDI-Berichte 1960, Düsseldorf, pp 19–34
- Helander MG, Zhang L (1997) Field studies of comfort and discomfort in sitting. Ergonomics 40(9):895–915
- Herzberg HTE (1958) Seat Comfort. Annotated bibliography of applied physical anthropology in human engineering. WADC Technical Report 56-30, pp 297–300
- Kompass K, Huber W (2006) Wie weit darf Fahrerassistenz gehen? Advanced driver assistance – how far should they go? VDA Technischer Kongress 22.03-22.03 2006
- Kompass K, Reichart G (2006) Freude am Fahren zwischen Selbstbestimmung und autonomer Technik AAET 2006 – Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel, GZVB Braunschweig
- Knapp A et al (2009) Code of practice for the design and evaluation of ADAS. http://www.acea.be/images/uploads/files/20090831_Code_of_Practice_ADAS.pdf, Accessed 5 May 2010
- Naab K (2000) Automatisierung im Straßenverkehr automatisierungstechnik. Oldenbourg Wissenschaftsverlag GmbH, München
- Naab K, Reichart G (1998) Grundlagen der Fahrerassistenz und Anforderungen aus Nutzersicht. Seminar “Driver assistance systems”, Haus der Technik, Essen, 16./17.11.1998
- Peden M, Scurfield R et al (eds) (2004) World report on road traffic injury prevention. WHO, Geneva

- Rasmussen J (1983) Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models pullman. IEEE Trans Syst Man Cybern 13(3):257–266
- Reichart G (2001) Menschliche zuverlässigkeit beim Führen von Kraftfahrzeugen. VDI Verlag, Düsseldorf
- Wenzel HG (1993) Klima. In: Schmidtke H (ed) Ergonomie. Carl Hanser Verlag, München
- WHO (2009) Global status report on road safety: time for action. Geneva, World Health Organization www.who.int/violence_injury_prevention/road_safety_status/2009, Accessed 21 July 2011

24 Adaptive Cruise Control

Hermann Winner

Technische Universität Darmstadt, Fachgebiet Fahrzeugtechnik,
Darmstadt, Germany

1	<i>Introduction</i>	615
2	<i>Requirements</i>	617
2.1	Functional Requirements for Standard-ACC Pursuant to ISO 15622	617
2.2	Additional Functional Requirements for FSR-ACC Pursuant to ISO 22179	618
3	<i>System Structure</i>	619
4	<i>ACC State Management and Human–Machine Interface</i>	620
4.1	System States and State Transitions	620
4.2	Control Elements	622
4.3	Display Elements	623
5	<i>Target Object Detection for ACC</i>	625
5.1	Requirements of the Environmental Sensors	625
5.2	Measurement Ranges and Accuracies	625
5.2.1	Distance	625
5.2.2	Relative Speed	626
5.2.3	Lateral Detection Area for Standard ACC Function	627
5.2.4	Lateral Detection Range for FSRA	628
5.2.5	Vertical Detection Range	629
5.2.6	Multi-target Capability	630
6	<i>Target Selection</i>	630
6.1	Determination of the Path Curvature	630
6.2	Path Prediction	632
6.3	Driving Corridor	633
6.4	Further Criteria for Target Selection	636
6.5	Target Selection Limits	636
7	<i>Vehicle-Following Control</i>	638
7.1	Basic Observations with Respect to Vehicle-Following Control	638

8	<i>Target-Loss Strategies and Curve Control</i>	640
8.1	Approach Strategies	641
8.2	Reaction to Static Objects	642
9	<i>Longitudinal Control and Actuators</i>	642
9.1	Basic Structure and Coordination of Actuators	642
9.2	Brake	643
9.2.1	Actuator Dynamics	644
9.2.2	Control Comfort	644
9.2.3	Feedback Information	645
9.3	Drive	645
9.3.1	Engine Control (Control Range, Actuator Dynamics, Steps/Accuracy, Feedback Information [Loss Torque of Ancillary Units])	646
9.3.2	Transmission Control	647
10	<i>Use and Safety Philosophy</i>	648
10.1	Transparency of the Function	648
10.2	System Limits	648
11	<i>Safety Concept</i>	650
12	<i>Users and Acceptance Studies</i>	650
12.1	Acceptance	650
12.2	Use	651
12.3	Driver-Control-Take-Over Situations	652
12.4	Comfort Assessment	654
13	<i>Conclusion and Outlook</i>	654
13.1	Current Developments	654
13.2	Function Enhancements	654

Abstract: Adaptive Cruise Control (ACC) has reached a new quality in driver assistance. For the first time, a large part of the driver's tasks can be assigned to an automatic system and the driver relieved to a substantial degree. Based on Cruise Control, ACC adjusts the vehicle speed to the surrounding traffic. It accelerates and decelerates automatically when a preceding vehicle is traveling at less than the speed desired by the driver.

ACC is a key functional innovation and represents a new system architecture with a high degree of function distribution. The different operating modes and system states are described along with function limits and transition conditions.

From the many elements of this overall function, target selection and longitudinal control are addressed in detail because of the special challenges they present. Target selection is based on the actual road curvature being determined by the ESC sensor signals that describe the driving dynamics, of which several options are assessed. Predicting and selecting a suitably shaped corridor is explained using an example. Major sources of error for the individual steps, their severity, and possible countermeasures are described.

The prerequisite for vehicle-following-distance control is the selection of a target. An example shows that the basic control principle is simple, but it conflicts with comfort and convoy stability. Details of additional control functions in curve situations and approaches are provided.

The driver perspective is addressed in terms of control and display functions and in terms of satisfaction as ascertained by use and acceptance studies, also taking into account an extended driver familiarization phase.

1 Introduction

Adaptive Cruise Control, abbreviated to ACC, describes a method of vehicle speed control which adapts to the traffic situation. Active cruise control, automatic distance control, automatic cruise control, or autonomous intelligent cruise control tend to be used as synonyms. Distronic and Automatic Distance Control (ADR) are registered trademarks.

The relevant international standards are ISO 15622 (Transport information and control systems – Adaptive Cruise Control systems – Performance requirements and test procedures) (ISO TC204/WG14 2002) and ISO 22179 (Intelligent transport systems – Full-Speed-Range Adaptive Cruise Control (FSRA) systems – Performance requirements and test procedures) (ISO TC204/WG14 2008), and, with the former describing the first functionality, often referred to as the standard ACC, while the second describes an extension of the functionality for the low-speed range, known as a full-speed-range ACC.

In ISO 15622, the ACC function is described as follows:

- An enhancement to conventional cruise control systems, which allows the subject vehicle to follow a forward vehicle at an appropriate distance by controlling the engine and/or power train and potentially the brake.

ACC is derived from the long-standing Cruise Control which is widely used in North America and Japan (abbreviated to CC). Its role is to control a desired speed v_{set} set by the driver, and it is included as part of the ACC function (🔍 Fig. 24.1, top).

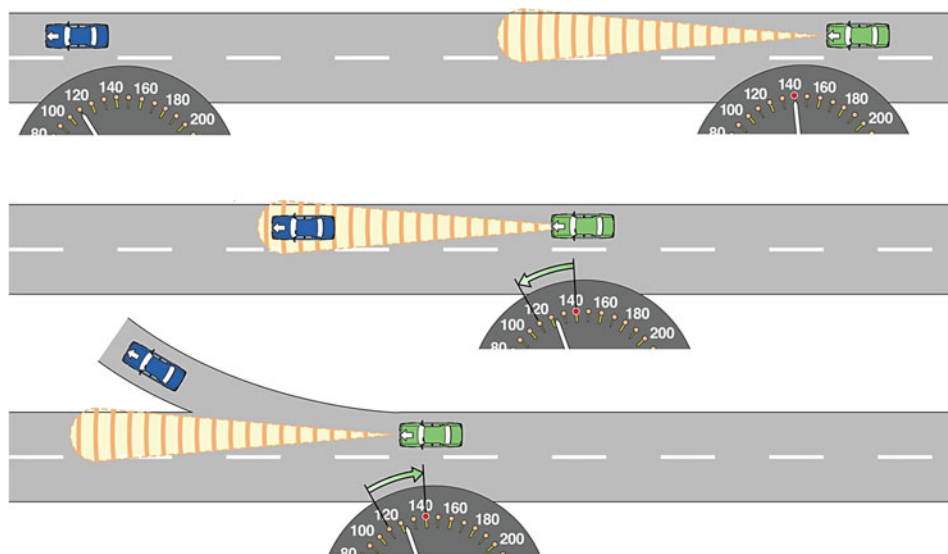
The main extension concerns adjusting the speed to the speed of the immediately preceding vehicle, here in addition to v_{to} (to: target object identified by ACC as target for control) (🔍 Fig. 24.1, center).

Although ISO 15622 leaves it open as to whether the brake is used for the control, the application of the brake to increase the deceleration has become established as a de facto standard. The appropriate distance mentioned in this standard is determined by τ , the time gap that is often colloquially referred to as distance in seconds. It is defined as:

- Time gap τ : "Time interval for traveling a distance, which is the clearance d between consecutive vehicles. Time gap is related to vehicle speed v and clearance d by: $\tau = d/v$."

The use of the temporal rather than the spatial reference follows the basic idea that to prevent rear collision, it is sufficient to have a distance which is in accordance with the reaction time, assuming the same deceleration capability for both the preceding and the subject vehicle. Therefore, in the presence of a preceding vehicle which is moving slower than one's own desired speed, the control task of the ACC is to adapt one's own speed to that of the preceding vehicle to obtain compliance with a clearance that ensures a constant reaction time.

However, as soon as the target leaves the immediate driving corridor and no other vehicle is designated as the target, ACC restores the reference speed without further action by the driver (🔍 Fig. 24.1).



■ Fig. 24.1

Situation-adapted change from free driving to following and back (Source: BOSCH)

2 Requirements


2.1 Functional Requirements for Standard-ACC Pursuant to ISO 15622

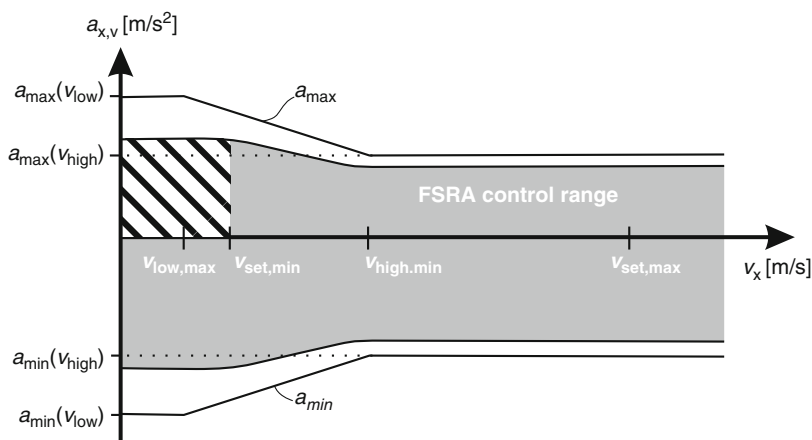
The function definitions in [Sect. 1](#) give rise to the following functional requirements:

1. Under free cruising conditions:
 - (a) Constant speed control and high control comfort, i.e., minimal longitudinal jerk and no swinging but high control quality (with no obvious deviation from the set speed)
 - (b) Cruise control with brake intervention in case of a lowered desired speed or incline travel
2. When following another vehicle:
 - (a) Vehicle-following control with swinging damping adjustment to the speed of the vehicle ahead so that the speed fluctuations are not copied
 - (b) Time-gap control to maintain the set time gap τ_{set}
 - (c) Control with the dynamics expected by the driver
 - (d) Smooth “falling back” like the standard behavior of a driver in case of a cut-in
 - (e) Convoy stability of the control when following other ACC vehicles
 - (f) Adequate acceleration capability for dynamic following
 - (g) Ability to decelerate for the majority of pursuit driving situations (90%) in moving traffic
 - (h) Automatic target detection when approaching or cut-in or cut-out situations within a defined distance range, i.e., determination of a target-seeking corridor
3. When approaching:
 - (a) For slow approaches, prompt speed control to the desired distance
 - (b) For faster approaching, predictable deceleration course in order to facilitate an assessment by the driver of whether to intervene because of inadequate ACC deceleration
 - (c) If the vehicle has become closer than the desired clearance, “falling back” in a standard driving manner
4. Functional limits:
 - (a) No control at very low speeds, i.e., hand over to the driver while the speed is below a minimum speed (ISO 15622: below $v_{\text{low}} \leq 5$ m/s no positive acceleration)
 - (b) Minimum set speed $v_{\text{set,min}}$ above 7 m/s ($\rightarrow 30$ km/h or 20 m/h speedometer values)
 - (c) The time gap not below $\tau_{\text{min}} = 1$ s in the steady state
 - (d) Priority given to driver’s intervention, i.e., deactivation when brake pedal is depressed and override when accelerator pedal is depressed
 - (e) Driver to set desired speed v_{set} and desired time gap τ_{set}
 - (f) Appropriate handover in the event of system failure, in particular when this occurs during deceleration
 - (g) Acceleration within the limits of $a_{\text{min}} = -3.5$ m/s² to $a_{\text{max}} = 2.5$ m/s²

2.2 Additional Functional Requirements for FSR-ACC Pursuant to ISO 22179

In addition to the requirements for the standard ACC function, Full-Speed-Range ACC has the following additional requirements:

1. When following another vehicle:
 - (a) Control throughout the entire speed range down to 0 km/h, particularly in the creep speed range (with increased requirements for the coordination of drive train and brakes)
2. When stopping:
 - (a) Control of appropriate stopping distance (typ.: 2–5 m)
 - (b) Greater deceleration capability at low speeds (refer  Fig. 24.2)
 - (c) Safe stopping with service brake in active system mode
 - (d) In the case of system shutdown to a standstill without driver intervention, transition into a safe holding state without power supply is required
3. Functional limits:
 - (a) Above of $v_{\text{high,min}} = 20 \text{ m/s}$ an acceleration within the limits of $a_{\text{min}}(v_{\text{high}}) = -D_{\text{max}}(v_{\text{high}}) = -3.5 \text{ m/s}^2$ up to $a_{\text{max}}(v_{\text{high}}) = 2.0 \text{ m/s}^2$ is permitted.
 - (b) Below $v_{\text{low,max}} = 5 \text{ m/s}$ acceleration within the limits of $a_{\text{min}}(v_{\text{low}}) = -D_{\text{max}}(v_{\text{low}}) = -5.0 \text{ m/s}^2$ up to $a_{\text{max}}(v_{\text{low}}) = 4.0 \text{ m/s}^2$.
 - (c) Between $v_{\text{low,max}}$ (5 m/s) and $v_{\text{high,min}}$ (20 m/s) the acceleration shall be between the speed-dependent limits of $a_{\text{min}}(v) = -D_{\text{max}}(v) = -5.5 \text{ m/s}^2 + (v/10 \text{ s})$ up to $a_{\text{max}}(v) = 4.67 \text{ m/s}^2 - (2v/15 \text{ s})$.
 - (d) The rate of deceleration γ below 5 m/s shall not exceed the jerk limit of $\gamma_{\text{max}}(v_{\text{low}}) = 5 \text{ m/s}^3$ and above 20 m/s of $\gamma_{\text{max}}(v_{\text{high}}) = 2.5 \text{ m/s}^3$. Between these parameters, the limit depends on the speed: $\gamma_{\text{max}}(v) = 5.83 \text{ m/s}^3 - (1v/6 \text{ s})$.



■ Fig. 24.2

Functional limits of FSR-ACC according to ISO 22179

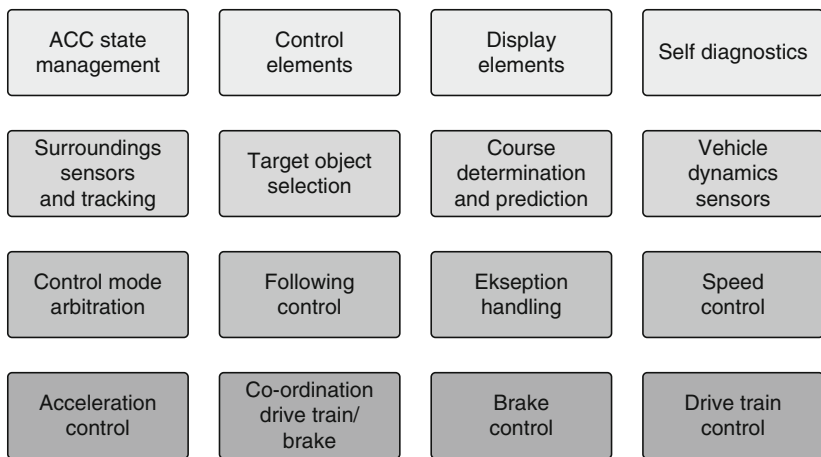
3 System Structure

The various tasks of ACC relate to the modules shown in the diagram in Fig. 24.3. The modules can in turn be subdivided, and different hardware units can be assigned. The information interfaces between the modules can vary significantly. This affects both the physical content and the data rate and bit representation.

ACC was the first system to influence vehicle dynamics which, as a distributed system, would lose its core functionality in the event of the failure of a peripheral system component. The other vehicle speed control functions (cruise control CC) with no adaptive capability can still be provided if all the necessary systems for speed control, i.e., engine, brakes, indicators, and controls are available.

If multiple sensors are used for ACC, further degradation steps are possible. For example, if short-range radar is not available, but the long-range sensor is, (reasons might include heavy snow because the long-range sensor, unlike the short-range sensors, is equipped with a lens heater, or because the short-range sensors must be turned off in the vicinity of radio astronomy stations), then above the minimum set speed the system function will switch to standard ACC. Only below the lower speed threshold is the system shut down in conjunction with a driver warning. Thus, the FSRA system in this speed range has no loss of availability compared to the standard ACC system.

With degradation of the systems it has to be ensured that the resulting changes in the system properties appear plausible to the drivers at all times and that they can adjust their thinking and thus the predictable system responses to the changed situation. Corresponding display options are to be provided. If, due to the limits of the operation and display concept, a clear differentiation is not guaranteed, it must be considered whether to prioritize increased availability or unambiguous system functionality.



■ Fig. 24.3
Function modules of ACC systems

4 ACC State Management and Human–Machine Interface

4.1 System States and State Transitions

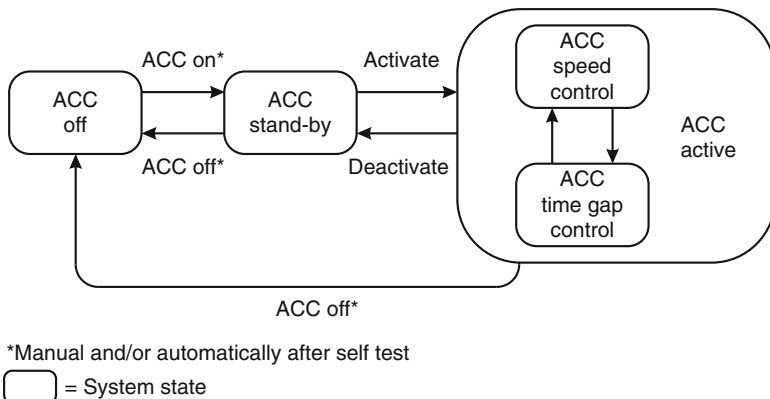
The system states of the standard ACC are illustrated in [Fig. 24.4](#). The On state is the ACC-off state, which, after a successful automatic function test, can be switched automatically to the ACC *stand-by* state or directly by the driver with the main switch. This stand-by state, insofar as the defined criteria for activation (see [Table 24.1](#)) are met, enables activation of the ACC *active* state.

If the ACC has been successfully activated, two major control states exist: *Speed control* in situations of free cruising and *ACC time gap control* for following a vehicle ahead that is traveling at a lower speed than the set speed v_{Set} . If this is not the case, the speed is adjusted to the desired speed v_{Set} . Transition between these states usually takes place automatically without driver intervention, purely by the detection of a target object and its distance and speed by the forward-ACC-sensor, as illustrated in [Fig. 24.4](#).

Deactivation, i.e., transition from ACC *active* to ACC *stand-by*, is usually initiated by actuation of the brake pedal or intentionally switching off via the control button. The various systems on the market have even more deactivation criteria which are shown in the right-hand column of [Table 24.1](#). Transition into the ACC-off state is effected when malfunctions are detected and by the main switch, if present. [Section 5](#) describes further control options and display functions.

For implementation of the ACC as a Full-Speed-Range-ACC, basically just one state is added: the *FSRA-hold* and its transitions. This is described in [Fig. 24.5](#).

The *FSRA-hold* state marks the holding of the vehicle at a standstill by the FSRA system. A transition from the *speed control* to the *hold* state would require a desired speed 0 km/h to be permitted. It makes sense to limit the minimum desired speed $v_{\text{set,min}}$ to a value >0 , e.g., 30 km/h.



■ Fig. 24.4
 States and transitions according to ISO 15622

■ Table 24.1

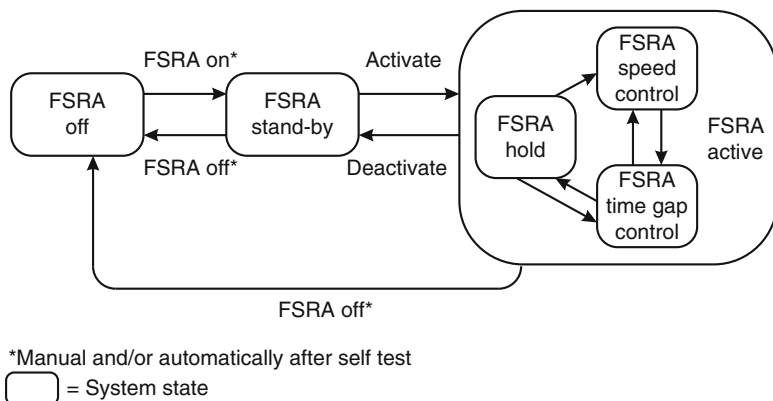
Activation and deactivation criteria (for activation all criteria have to be fulfilled, for deactivation just one)

Activation, when simultaneously....	Deactivation in the case of....
	Deactivation by control switch
	Driver brakes at $v > 0$
$v \geq v_{\text{set,min}}$	$v < v_{\text{min}}$ (only standard ACC)
Engine speed not significantly below idle speed	Engine speed significantly below idle speed
Forward gear engaged	Nonvalid gear (with automatic transmission: N-position) With manual transmission: longer periods of (>8 s) disengaged clutch or engaged clutch w/o engaged gear
ESP/DSC in full operation	ESP passive
Slip control not active	Slip control active for more than a given time (depends on the cause, e.g., 300 ms for yaw control, ca. 600–1,000 ms for traction control)
Parking brake released	Parking brake activated
No ACC system failure or sensor blindness	ACC system failure or sensor blindness
<i>Additionally for FSRA:</i>	
Driver's door closed	$v = 0$ AND at least two of three signals active: door open, no belt, seat not occupied
Driver seatbelt is fastened (using seat occupancy sensors)	
Brake pedal acting AND $v = 0$ AND target object detected	Note: At ($v = 0$ AND driver brakes) no deactivation
Target object detected AND $0 < v < v_{\text{set,min}}$	

In the *hold* state, some special features are observed. Even if it is possible to transfer the holding function to the driver with appropriately signaled instructions, it is good practice to keep the vehicle on, even by simply actuating the brake pedal. The system does not switch off, but the vehicle is kept safe from inadvertently rolling and so prevents critical system states. If the system is switched off intentionally, this must usually be done by a dual action, for example, by operation of the ACC-off switch while depressing the brake pedal.

For safety reasons and except for very short stops, the transition from a *hold* state into one of the two driving states is only allowed with driver's confirmation, as the current state of the sensors is not sufficiently reliable to detect all contingencies that may take place in the state.

Similarly, the presence of the driver is monitored as s/he can leave a stopped vehicle at any time. Upon detection of an intention to exit (e.g., open door, released belt, or no seat



■ Fig. 24.5

States and transitions for FSR-ACC according to ISO 22179

occupancy detection) a suitable system shutdown with a safe hold state is initiated, even in the event of power failure, for example, by activating an electromechanical parking brake. If this is not possible, the driver must be warned and/or the system switched off before the driver leaves the vehicle so that s/he is able to secure the vehicle against rolling away.

Once stopping is detected, the responsibility for safe holding is transferred to the ESP system. For a short time, the braking pressure must be increased to ensure sufficient lock and permanent holding without power, controlled by an electric parking brake (EPB).

4.2 Control Elements

The ACC control elements are used to implement the transitions from one state to another and adjust the preset control values, namely the desired speed and the desired time gap.

- Control element to switch from ACC-*off* to ACC-*stand-by* state. There are two options:
 - A switch which is activated only once and then remains permanently in the ON position.
 - A push button which activates the controls once per ignition.
- Control element for activation of the ACC system. This control is often also used for active control to increase the current set speed.
- Control element to reduce the current set speed.
- Control element to activate the ACC system, using the last set speed (Resume).
- Control element for setting the desired time gap. Here again, there are two fundamentally different power-up states:
 - A constant initial state with a default setting, which usually corresponds to a time gap of 1.5 to 2 s.
 - The last selected state, such as a mechanical lock.



■ Fig. 24.6

Control element for DISTRONIC PLUS (Mercedes-Benz W221) with seven functions

Often, the controls are arranged in groups or integrated into control levers, as the following examples illustrate.

The cruise control lever illustrated in ● Fig. 24.6 combines seven functions for the DISTRONIC PLUS of Mercedes-Benz vehicles. The actions 1 (up) and 5 (down) activate the ACC, initially adopting the current speed as the set speed. With further upward movements, the set speed is changed in small increments of 1 km/h, with large strokes in 10 km/h increments. With corresponding downward movements, the set speed is reduced in the same way. Movement 4 (toward the driver) also activates ACC, but it resumes at the previously used set speed (Resume function). The first time you activate the ACC, the current speed is adopted. Resume from standstill starts with this function. Movement direction 7 (forward) deactivates the ACC, while the movement 6 switches between cruise control and speed-limiter function. The operation of the speed-limiter function is analogous to the operation of cruise control/DISTRONIC PLUS. On activation, the LED lights up in the cruise control lever. Operating element 2 is turned to set the desired time gap. The last set rotational position is thus available for a new driving cycle, which can use the former setting.

4.3 Display Elements

Although most ACC states can be determined from the current control action, clear feedback of the states, especially during state transition, is also important for monitoring the system. However, the response of the desired speed and the desired time gap are also essential for user-friendly operation. We now differentiate between two types of display: permanent and situational. The latter appear only when a certain event occurs or for a certain time when the driver has operated a control element. The situational display has on the one hand the advantage that the display space can be shared with other situational display functions, and on the other, that attention can be focused better.

■ Table 24.2

Display functions for ACC

States	Type	I	T
Activation state	p	e	o
Relevant target object detected	p	i	o
Override by the driver	s	h	o
Go possible (FSRA only)	s	h	o
Transition autom. go → driver triggered go (FSRA only)	s	h	o
System settings			
Desired speed (set speed)	p	e	o
Desired time gap (set time gap)	p,s	i	o
Speed of preceding vehicle	p	h	o
Actual distance to preceding vehicle or deviation between set time gap to actual time gap	p	h	o
State transitions			
ACC off → ACC stand-by, if provided	p	e	o
Handover request when a system limit is reached	s	i	a + o
System shutdown	s	e	a + o
Below a critical distance or time-gap limit	s	i	a (+o)

A further distinction concerns the importance (I) of the display, distinguishing between the stages *essential*, *important*, and *helpful*. According to this classification, *essential* displays are found in all systems and important displays are found in *most*. Even while *helpful* displays are only implemented in some cars, they improve the intuitiveness of the system functions and give the driver detailed expectations and a better understanding of the system responses. The display of distance and relative speed of the newly detected object enables the driver to easily spot a false detection and assign the system responses plausibly.

As is often the case with control functions, the activation state and the set speed are also combined in the display. ● Table 24.2 shows the best-known display functions and the recommended display technology (T), where, in this case, initial differentiation is only between optical (o) or audio elements (a). Haptic elements are not considered since haptic display functions are not used for ACC apart from the inherent kinetic feedback of the handling.

The display of the Mercedes-Benz Distronic shown in ● Fig. 24.7 contains details about the activation state, depicting the subject vehicle (● 24.2), the set speed (marking in the speedometer crown), the desired distance (bar “under the road”), target detection (vehicle 4), indicating the position of the actual clearance, and the speed band which is bounded below by the speed of the vehicle ahead and above by the set speed. Not listed in the picture are the symbols for the take-over-request and the display when the driver overrides the ACC (Distronic passive).



■ Fig. 24.7
Displays of Distronic Plus (Mercedes-Benz W221)

5 Target Object Detection for ACC

5.1 Requirements of the Environmental Sensors

The ACC functionality succeeds or fails with the detection of the relevant target vehicle, which is the basis for the control. The first prerequisite is a set of surrounding sensors which are necessary to detect and then to decide, of the vehicles in the relevant area, whether and which of the detected objects is to be selected as the target object. Radar and lidar are successfully used as surroundings sensor technologies. The requirements listed below apply equally to both.

5.2 Measurement Ranges and Accuracies

5.2.1 Distance

According to subdivisions defined for ISO 15622 (2002), the standard ACC function requires that objects are detected from the minimum detection distance $d_{\min 0} = \text{MAX}(2 \text{ m}, (0.25 \text{ s} \cdot v_{\text{low}}))$ and, also, the distance determined from $d_{\min 1} = \tau_{\min}(v_{\text{low}}) \cdot v_{\text{low}}$ is the smallest time gap at the smallest allowed ACC operation speed. As the time gap is increased at low speeds, $d_{\min 1}$ is about 10 m. There is no need for distance measurement below this distance because the ACC control will always decelerate in this situation or in any case the driver is asked to take over at speeds of below v_{low} . If speed falls below $d_{\min 0}$, it can be assumed that a control process will be interrupted before reaching such a small distance from the driver. The same applies in the case of a cut-in in close proximity to the subject vehicle, for which drivers will not rely on the ACC function but will resolve the situation through their own brake operation.

The maximum distance d_{\max} required must, of course, enable control with the maximum target distance, i.e., the distance for setting the largest time gap at the

■ Table 24.3

Distance requirements of typical setups

$\tau_{\text{set,min}} (v_{\text{low}}) = 2 \text{ s}$	$v_{\text{low}} = 5 \text{ m/s (=18)}$	$d_{\text{min0}} = 2 \text{ m}$
		$d_{\text{min1}} = 10 \text{ m}$
$\tau_{\text{set,max}} = 2 \text{ s}$	$v_{\text{set,max}} = 50 \text{ m/s (=180 km/h)}$	$d_{\text{max}} = 100 \text{ m}$

maximum setting speed $v_{\text{set,max}}$. A control margin is usually reserved for the comfort of the system control. Since a set time gap of at least $\geq 1.5 \text{ s}$ is required, the maximum time gap can only be reduced as far as this limit.

The above requirements (► Table 24.3) are minimum requirements and apply only to stationary pursuit. A greater distance is desirable for an approach, particularly if the speed difference is considerable. As shown later, the target selection is more difficult the greater the distances involved, consequently, in many cases a deceleration reaction is experienced as negative at a distance greater than 120 m even if the target selection is working without error. This is particularly the case if you intend to overtake the target vehicle. Overtaking is impeded by the ACC deceleration reaction before the lane change has started.

In practice (Winner and Olbrich 1998), a restriction of the reaction range has been useful. Particularly in the lower and middle speed range, there is no benefit from reacting throughout the entire range since objects have no effect on one's own vehicle at a great distance.

No high demands are made on the accuracy of distance measurement as the system responds only weakly to distance variations.

5.2.2 Relative Speed

The accuracy of the relative velocity must fulfill far higher requirements than that for distance. Any deviation of the relative velocity leads to a change of acceleration (see ► Sect. 7). A static offset leads to a steady deviation of the distance, with an offset of 1 m/s leading to an approximately 5-m distance deviation. Fluctuations in the speed of $v_{\text{rel,err}} = 0.25 \text{ m/s}$ (rms in the 0.1 to 2 Hz band) are still accepted as the resulting subsequent acceleration fluctuations remain below the driver's sensitivity. While filtering the speed signal can reduce the fluctuations effectively, an excessive delay has to be avoided as otherwise the control quality is adversely affected. As a guideline, a maximum delay time of 0.25 s can be used, wherein for stable control with the smallest time gap of $\tau_{\text{min}} = 1 \text{ s}$, 0.75 s remains for the control time constant and the actuator delay.

Relative errors $\varepsilon_{\text{vrel}}$ of the relative speed up to 5% are largely unproblematic for vehicle-following control, since the consecutive acceleration control with the control systems for brake and drive train produce similarly large deviations and thus the relative distortions of the control set point caused by the speed error are hardly noticeable.

Greater challenges for the accuracy of the relative velocity are posed by the classification of objects, whether they are moving in the same direction, at a standstill or moving in the

opposite direction. For this classification, tolerances must be smaller than 2 m/s and 3% of v_{rel} . The relative error can also be calibrated with stationary objects because they are measured much more frequently and are thus seen as an accumulation in a statistical measurement. This allows even those errors to be compensated that result from vehicle speed determination based on the dynamic wheel radius, whose accuracy is usually limited to 2%.

5.2.3 Lateral Detection Area for Standard ACC Function

The requirements for the lateral detection range are derived from the initial assumptions:

- τ_{max} , the maximum time gap for following-distance control
- $a_{y\text{max}}$, the maximum assumed lateral acceleration for cornering
- R_{min} , the smallest curve radius specified for the ACC function

For a given curve radius $R \geq R_{\text{min}}$, the maximum cornering speed can be determined by the maximum lateral acceleration. If this is multiplied by the time gap τ_{max} , we obtain the required maximum range $d_{\text{max}}(R)$. The offset of the curve line y_{max} at d_{max} (● Fig. 24.8), however, is independent of the curve radius and speed:

$$y_{\text{max}} = \frac{(\tau_{\text{max}}^2)}{2} \cdot a_{y\text{max}} \quad (24.1)$$

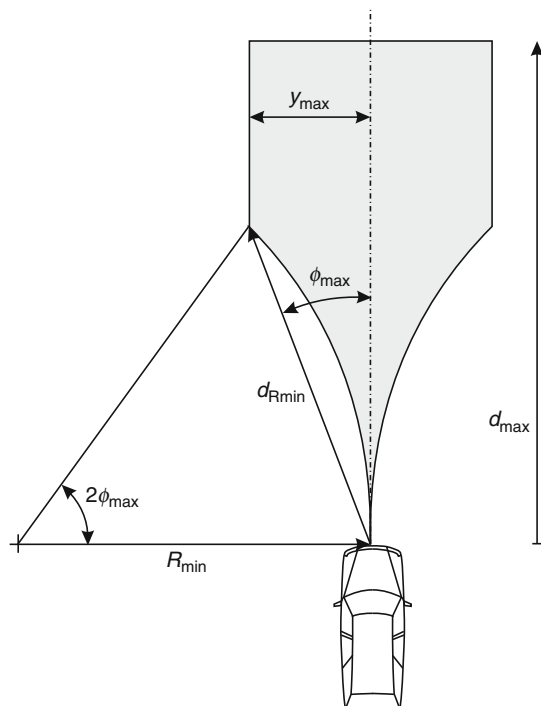
The maximum azimuth angle ϕ_{max} is determined by the ratio of the maximum offset y_{max} and the maximum range d_{max} at $R = R_{\text{min}}$:

$$d_{R\text{min}} = d_{\text{max}}(R_{\text{min}}) = \tau_{\text{max}} \sqrt{a_{y\text{max}} \cdot R_{\text{min}}} \quad (24.2)$$

$$\phi_{\text{max}} = \arcsin(y_{\text{max}}/d_{\text{max}}(R_{\text{min}})) \approx y_{\text{max}}/d_{\text{max}}(R_{\text{min}}) \quad (24.3)$$

Because of the observed driver behavior (e.g., see Mitschke et al. 1991) the underlying fundamental lateral acceleration is dependent on the driving speed. Implicitly, this results in a dependence on the curve radius as tighter bends are traversed at lower speeds. This is reflected by the different values for the standard curve classes defined in ISO 15622. Thus, $a_{y\text{max}} = 2.0 \text{ m/s}^2$ is adopted for $R_{\text{min}} = 500 \text{ m}$ and $a_{y\text{max}} = 2.3 \text{ m/s}^2$ for $R_{\text{min}} = 250$ and $R_{\text{min}} = 125 \text{ m}$, respectively. In ● Fig. 24.9, for a maximum time gap of $\tau_{\text{max}} = 2 \text{ s}$, the necessary (unilateral) angle ϕ_{max} is illustrated for three different maximum lateral acceleration assumptions. Despite this highly idealized real cornering view, measurements from the field (Winner and Luh 2007; Luh 2007) show that the formula above and the assumptions can be used to determine the requirements for the opening angle curve for a given capability. The two empirical values refer to the curve radius at which half of the following-distance-controlled runs occurred without a target loss.

Another outcome of the investigations (Winner and Luh 2007; Luh 2007) showed that with an opening angle of $\Delta\phi_{\text{max}} = 16^\circ (\pm 8^\circ)$, both subjectively and objectively, the standard ACC function is covered to a sufficient extent and a further increase of the azimuth angle range results in less improvement of the standard ACC function as long as



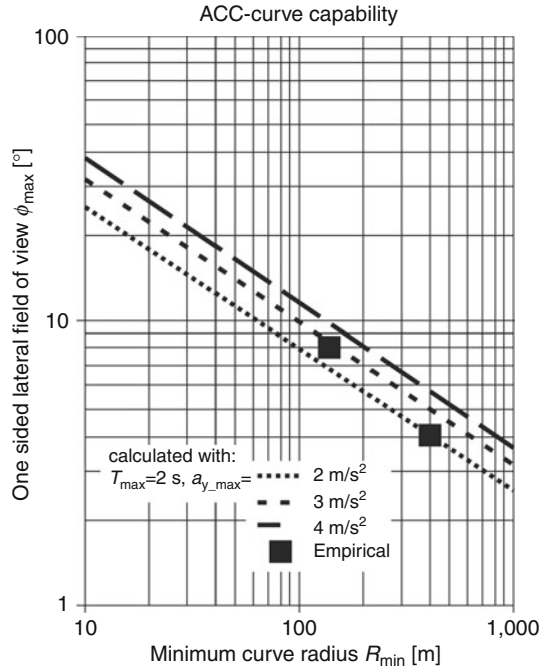
■ Fig. 24.8

Required detection range (in azimuth angle) depending on the curve radius at constant lateral acceleration and time gap

the detection of cut-in vehicles by the dynamics of target selection is specified (see ▶ Sect. 6). As evident in the subsequent consideration of the overall error, a small azimuth alignment error leads to significant functional impairment. Since the tolerance limit depends on many individual factors, particularly on back-scattering properties of the objects, no definite value can be given. Static error of 0.25° should be avoided, while dynamic noise-like errors should be smoothed with filters and may amount to 0.5° without serious detrimental effects on the system function.

5.2.4 Lateral Detection Range for FSRA

Hundred percent coverage of the area directly in front of the vehicle is aimed at in order to enable an automatic go function. As this is difficult to realize in practice, the minimum requirements are much lower in the emerging FSRA ISO standard 22179 and can be met with a centrally positioned sensor with an opening angle $\Delta\phi_{\max} = 16^\circ (\pm 8^\circ)$. One such FSRA is the Audi Q7's ACC Plus. Standing starts in such systems are therefore limited to a corresponding driver-enabling function.



■ Fig. 24.9

Required azimuth angle range depending on assumed lateral acceleration and time gap.
Lines: theoretical path; *dots:* experimental results for two angle ranges

An excess of $\pm 8^\circ$ coverage of the area in front of the vehicle up to about 10–20 m is required for close and staggered following-distance control at low speeds. This most likely occurs in congested conditions when not driving directly behind the vehicle ahead, but also helps to achieve a better view. Even if the target object changes lanes slowly, coverage will become increasingly difficult if one's own required driving corridor is not free. A sensor with a too narrow azimuth angle range will lose the target object, although it could not be passed without colliding. So the driver has to intervene in these cases.

From the minimum distances which can be typically expected for the cut-in, i.e., about 2–4 m at very low speeds, coverage of neighboring lanes also makes sense (at least up to half the width) to ensure an early detection of cut-in vehicles. Reliable angle determination is very important as only by calculating the lateral movement from the angle values can ACC respond to cut-in vehicles.

5.2.5 Vertical Detection Range

Vertical coverage requires the detection of all relevant objects for ACC (trucks, cars, motorcycles). Since the objects are not very high off the ground or are lower than the

normal sensor installation heights, only the slope parameters and static and dynamic changes within the pitch utilized by ACC dynamics are considered. In practice, requirements of $\Delta\vartheta_{\max} = 3^\circ (\pm 1.5^\circ)$ have resulted.


Incorrect elevation angles usually have only a small negative effect as the elevation is rarely used as a measured variable, for example, 2D-scanning lidar scans the environment in several superimposed horizontal lines. However, in the case of radar sensors, changes of the antenna pattern with deviating elevation angles of greater than 0 can be expected. Furthermore, it is necessary to prevent the elevation of the available area from being reduced by a misalignment to the extent that the above requirement is no longer guaranteed.

5.2.6 Multi-target Capability

As several objects may be present in the sensor field, a multi-target capability is very important. This particularly means the ability to differentiate between relevant objects in the driving corridor and irrelevant objects, such as those on the adjacent lane. This differentiation can be achieved by a high degree of differentiation of at least one measured variable (distance, relative velocity, or azimuth angle). However, the requirement for a high degree of differentiation must not be at the expense of association problems in which the objects are identified repeatedly as new objects.

6 Target Selection

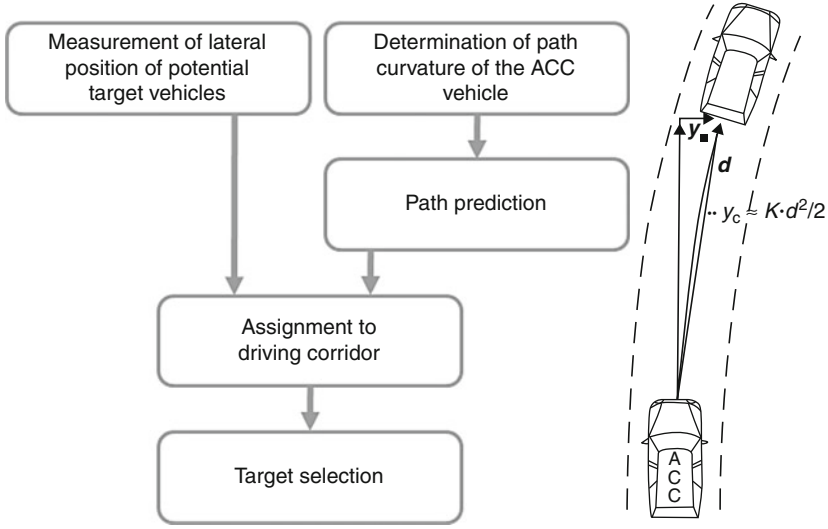
Target selection is of very great importance to the quality of the ACC as both relevant objects can be overlooked (false negative) and false targets may be selected. In both cases, the user's expectations of this system are not fulfilled.

The following error analysis is based on the need for target selection, as shown in the steps in  Fig. 24.10, left. The measurement of the lateral position $Y_{U,i}$ of the object i is carried out by the ACC sensor with an uncertainty of $\varepsilon_Y \approx \varepsilon_\phi \cdot r$, which results from the inaccuracy of the angle determination ε_ϕ .

6.1 Determination of the Path Curvature

The curvature κ describes the change of direction of a vehicle as a function of the distance traveled. The constant part of a curve is the reciprocal of the curve radius $R = 1/\kappa$. The curvature of the vehicle trajectory can be determined by various onboard sensors and assumes for all calculations that they are used outside dynamic vehicle limits. So they are not valid for skidding situations or in the presence of significant wheel slip.

- Curvature calculated from the steering wheel angle



■ Fig. 24.10

Left: Steps for target selection; right: definition of the variables

In order to calculate the curvature κ_s from the steering wheel angle δ_H , three vehicle parameters are required: the steering ratio i_{sg} , wheelbase ℓ , and the characteristic speed v_{char} derived from the under-steering behavior in the linear dynamics range, characterized, therefore, at low lateral accelerations. Under typical ACC conditions of very good approximation, κ_s could be determined according to:

$$\kappa_s = \delta_H / [(i_{sg}\ell)(1 + v_x^2/v_{char}^2)] \quad (24.4)$$

- Curvature calculated from the yaw rate

To calculate the curvature κ_ψ from the yaw rate, the driving speed v_x is required and the slip angle rate is disregarded:

$$\kappa_\psi = \dot{\psi} / v_x \quad (24.5)$$

- Curvature calculated from the lateral acceleration

The calculation of the curvature κ_{ay} from the lateral acceleration a_y also uses the driving speed v_x :

$$\kappa_{ay} = a_y / v_x^2 \quad (24.6)$$

- Curvature calculated from the wheel speeds

For the curvature κ_v from the wheel speeds, the relative difference of the wheel speeds $\Delta v/v_x$ and the width of the track b are required. In order to minimize driving influences, the difference $\Delta v = (v_l - v_r)$ and driving speed $v_x = (v_l + v_r)/2$ are determined by the speed of the non-driven axle.

$$\kappa_v = \Delta v / (v_x b) \quad (24.7)$$

■ Table 24.4

Comparison of the different approaches for curvature determination

	κ_s	κ_{lf}	κ_{ay}	κ_v
Robustness against crosswind	— —	+	+	+
Robustness against lateral road inclination	— —	+	— —	+
Robustness against wheel radius tolerances	o	+	+	—
Sensitivity at low speeds	++	o	— —	—
Sensitivity at high speeds	—	o	++	—
Offset drift	+	— —	— —	+

Although all these methods can be used for the determination of curvature, they all have different strengths in different operating conditions. They differ particularly in crosswind, lateral road inclination, wheel radius tolerances, and, in terms of sensitivity, in different speed ranges.

As ❷ Table 24.4 shows, the curvature of the yaw rate is best. However, a further improvement in signal quality can be achieved if some or all signals are used for mutual comparison. This is especially possible because the ACC vehicle is equipped with ESP and, therefore, all the above sensors are part of the system. At standstill, there is offset adjustment of the yaw rate, but this requires a halt phase that does not occur when driving on highways with no traffic jams. In this case, statistical averaging methods can be used as the average of the yaw rate sensor supplies the offset over long distances.

6.2 Path Prediction

To predict the future path the system needs to know the (future) path of the carriageway and the future driving lane choice of the ACC vehicle and also, in fact, those of the potential target vehicles. Since this information is not available without image processing or vehicle-to-vehicle communication, working hypotheses are used, which employ simplified assumptions.

One simple hypothesis is the assumption that the current curvature will be retained. This basic hypothesis continues to be used until further information is available. This approach disregards entries and exits to bends, changes in the lane markings, and also drivers' steering errors. If past lane-marking assignments are available, the hypothesis that the objects and the ACC vehicle will remain within their lanes will be used. However, this is invalidated if objects cut in or out or if the driver changes lane. Nor does it help for the initial assignment.

The compromise approach is to delay the object data by half of the time gap and to allocate it on the basis of the then-current path curvature. This is very robust at the start and end of curves because, due to the delay, the curve of the road between the objects and

the ACC vehicle is used and, thereby, good assignment is permitted even when curvatures change. This method does not, however, constitute a replacement for the first one in the case of initial assignment.

Additional options for path prediction are offered by GPS navigation in conjunction with a digital map and the curvature information stored therein. Unfortunately, such maps are never up-to-date and roadworks are not marked. The method which uses static objects at the roadside to determine curvatures is also only partially useful in the absence of such objects, but is presumably nevertheless included in most ACC-path-prediction algorithms. The lateral movement of vehicles in front can also be used to improve path prediction because, in most cases, this is an early indication of an impending bend in the road.

The use of lane marking information from the processing of camera images is obviously very promising. However, an improvement at distances of over 100 m cannot really be expected today since today's standard camera pixels correspond to around 0.05° ; a value which, at around 120 m, corresponds to a width of 10 cm and is therefore barely adequate for lane-marking detection. In addition, image-based path prediction outside the headlight beam is impossible in darkness, especially when roads are wet.

The aforementioned algorithms are used in different ways and to different degrees by different manufacturers, but, overall, they deliver as a starting value a predicted trajectory κ_{pred} . The trajectory can then be extrapolated depending on the distance. Instead of the circular function, a parabolic approximation is sufficient in the case of the normal opening angle:

$$y_{c,U} = \frac{\kappa_{\text{pred}}}{2} d^2 \quad (24.8)$$

The cross-track error values $y_{i,U}$ of the object detected by the ACC sensor can then be related to this trajectory resulting in the relative offset:

$$\Delta y_{i,c} = y_{i,U} - y_{c,U} \quad (24.9)$$

Errors of the predicted curvature therefore increase fourfold with the object distance d . At high speeds ($v_x \geq 150$ km/h), a curvature error κ_{Err} of less than 10^{-4} m is acceptable, but it will cause an error at 100 m of $\Delta y_{i,c,\text{err}}(100 \text{ m}, 150 \text{ km/h}) \approx 0.5$ m. At 140 m, the error doubles due to the quadratic propagation. At low speeds (≈ 50 km/h), the curvature error pursuant to (32.5) is around three times higher. The distance for $\Delta y_{i,c,\text{err}}(57 \text{ m}, 50 \text{ km/h}) \approx 0.5$ m is just 57 m from which is derived a reduction of the maximum target-selection distance at low speeds.

6.3 Driving Corridor

The driving corridor is a term frequently employed by experts for the corridor used for the ACC target selection. In its simplest form, it is determined by a width b_{corr} (not dependent on distance) with the predicated path as a center line (➊ 24.8). Initially, one might equate

the driving-corridor width with the lane width b_{lane} . However, this assumption has been found to be inappropriate.

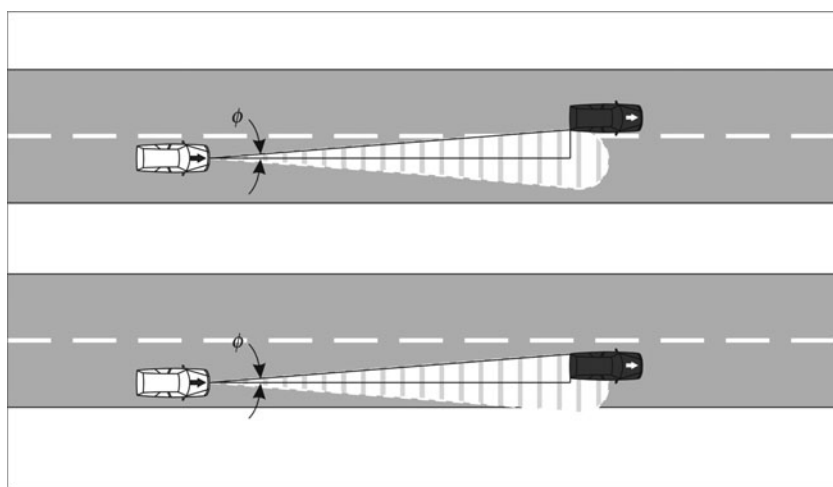
The example in [Fig. 24.11](#) shows that there is an area in which clear assignment is impossible based on the measured lateral position alone.

Since it cannot be assumed, however, that the measured lateral position of the object corresponds with the center of the object, both the right-hand and the left-hand object edges must be taken into account. Vehicles traveling off-center present further uncertainty of assignment, both of the ACC vehicle and of the potential target vehicle. Therefore, assignment to the actual lane is only certain if the measured lateral position (without errors) lies within ± 1.2 m of the predicted path center (without errors). The assignment of the object to the neighboring lane is only certain if its position is at least at least 2.3 m from the path center. The values relate to a lane width of 3.5 m.

Some misrecognition must be expected even in a lane width of 3.5 m, while in a narrower driving corridor target losses can be expected (Luh 2007).

Three measures are employed to improve target selection: a variable driving-corridor width depending on the type of road, an approximate driving-corridor contour, and a local and temporal hysteresis function for the target selection.

Two pieces of information are important for a variable driving-corridor width: Are there neighboring lanes to the left or right? If not, the selection area can be very wide on the respective side of the driving corridor (around 2 m to the respective side, i.e., 4 m if there is no driving lane in the same driving direction on either side). Information about the presence of neighboring lanes can be obtained from the observation of static targets at the roadside and by oncoming vehicles, whereby changes, e.g., broadening to two lanes in a single direction, can only be detected with a time delay. If a neighboring lane is detected, e.g., by the observation of vehicles in the same direction with a lateral position outside



■ Fig. 24.11

Example of differing assignments despite equivalent relevant data

one's own lane, a statistical observation of the lateral positions can be used to adjust the driving corridor so that roadworks involving narrower lanes can be negotiated.

A further measure is local hysteresis, which means that an object marked as a control object is assigned a wider driving corridor than all other objects. Typical differences are around 1 m, i.e., around 50 cm on either side. This prevents misrecognition of objects in the neighboring lanes particularly during changing conditions (entering and exiting bends, uneven steering) while nevertheless keeping the target object stable in these situations.

Temporal hysteresis is also used, as shown in [Fig. 24.12](#). As compared with assignment reliability (LP, lane probability), target plausibility (PLA) increases in the case of positive LP. Above an upper threshold (in this case, 0.4), the object becomes the target object unless other criteria suggest otherwise. The target plausibility may increase to a maximum value (in this case, 1) and decrease based on two options: In the case of the absence of detection (no signal) and if allocated to the neighboring lane (negative LP). Once below the lower threshold (in this case, 0.2), the object loses the characteristic of being able to be selected as the target object.

The assignment measurement LP can be approximately mapped as shown in [Fig. 24.13](#). The further away the object is, the less clear is the transition between the

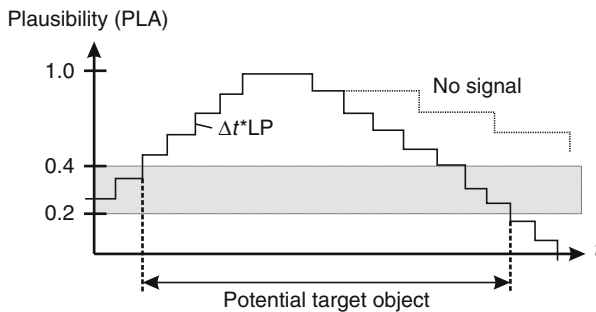


Fig. 24.12
Formation of target plausibility (illustration from Luh 2007)

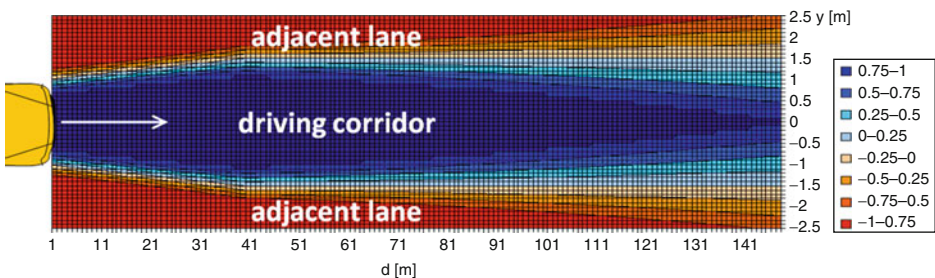


Fig. 24.13
Approximate corridor contour to avoid assignment errors

lane assignments. Therefore, account must be taken of the fact that errors of location determination and path prediction increase with distance. In addition, other estimated uncertainties can dynamically restrict the core range, e.g., a major bend in the road.

6.4 Further Criteria for Target Selection

It can make sense to use other criteria in addition to the lane assignment. The most significant criterion for target selection is the object speed. Oncoming vehicles are completely ignored for control purposes. Nor are static objects selected as target objects, with the exception of those already detected as objects moving in the travel direction (so-called halted objects). These are relevant particularly for the Full-Speed-Range ACC function in the same way as objects traveling in the same direction are relevant. Permanently static objects are often used for other functions (see also ● Sect. 8.2) and, therefore, are subject to separate filters. For basic ACC functions, however, they play only a minor role.

Another simple but very effective approach is to limit the distance as a function of the travel speed. Thus, at a speed of 50 km/h, a reaction to targets which are more than 80 m away is neither necessary nor expedient since the danger of mistaken assignment greatly increases with distance. Empirical values suggest a distance value $d_{to,0} = 50$ m and an increase of $\tau_{to} = 2$ s.

$$d_{to,max} = d_{to,0} + v \cdot \tau_{to} \quad (24.10)$$

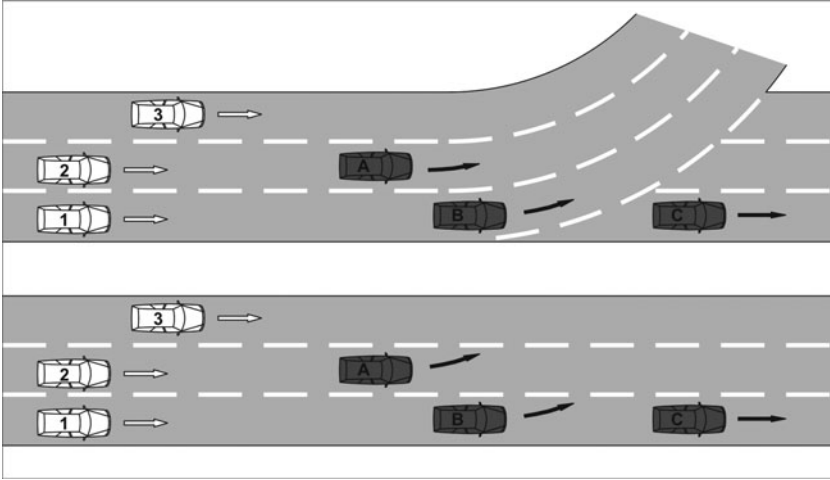
If several objects meet the criteria for a target object, the following decision-making criteria are considered individually or in combination:

- The smallest longitudinal distance
- The smallest distance to the path center (minimum $|\Delta y_C|$)
- The smallest set acceleration

The final criterion assumes a connection to the ACC control or a multi-target object interface, but improves the transition in the case of target objects cutting out.

6.5 Target Selection Limits

The approaches illustrated in the latter sections are highly effective and have reached a high quality level. There are, however, situation-dependent constraints which require explanation as in the following two examples. The vehicles in ● Fig. 24.14 move identically in the moment illustrated, assignment as the “correct” object, however, proceeds differently due to the differing progress of the road. Another example is the “overtaking dilemma” when approaching at high speed. For comfortable braking to follow a significantly slower moving vehicle, deceleration needs to start at a great distance



■ Fig. 24.14

Situation example for ambiguous target assignment (vehicle positions and movements are identical in both pictures)

away. On the other hand, the probability that the preceding vehicle must be overtaken is particularly high if the difference in speed is large. Early deceleration would considerably hinder the overtaking process. Since the overtaking process is rarely indicated more than 6 s before reaching the vehicle to be overtaken (Winner and Olbrich 1998), the dilemma exists between too early a response for unhindered overtaking and too late a response for a comfortable or even adequate approach.

Another parameter is the late detection of cutting-in vehicles. On the one hand, temporal and local hysteresis in the driving path assignment lead to a delayed reaction of around 2 s with respect to the moment the vehicle cutting in crosses the lane marking. Since drivers are aware of the cut-in even before the marking is crossed, due to the situation and the indication of the lane change by the direction indicators, the late reaction is once again a critical point for the user. The same phenomenon occurs with the detection of vehicles cutting out, although target approval is objectively correct once the neighboring lane has been fully entered.

Significant improvement of cutting-in and cutting-out detection on the part of ACC can only be achieved by situation classification. However, these kinds of “intelligent” functions affect the transparency of the ACC operation.

Improving the target selection in the case of lane changing on the part of the ACC vehicle can be achieved by interpreting the direction indicator, resulting in a shift of the driving path to the indicated direction. The digital map in combination with the search and detect function also enables an adaptive driving corridor function.

On average, modern ACC systems perform an incorrect assignment only once every hour; a value which is surprisingly small in the light of the many potential errors and one which is hard to improve.

7 Vehicle-Following Control

7.1 Basic Observations with Respect to Vehicle-Following Control

Although the ACC vehicle-following control is often described as a distance control, it is anything other than a means of difference-guided distance control. As a point of departure for further considerations, it is assumed that the controller output is manifested in direct vehicle acceleration with no time delay and, moreover, that the ACC vehicle follows the target vehicle at the set time gap τ_{set} . Disregarding vehicle lengths, it therefore follows that the ACC vehicle will reach the position of the target vehicle after a time gap of τ_{set} . If the ACC vehicle now echoes the position of the target vehicle with a time lag, the time gap will be retained irrespectively of speed. In the same way, the speed and acceleration of the preceding vehicle are imitated with a time lag. Thus, in the steady state, a simple control principle can be derived which even avoids feedback:

$$\ddot{x}_{i+1}(t) = \ddot{x}_i(t - \tau_{\text{set}}) \quad (24.11)$$

The index $i+1$ stands for the ACC vehicle in a convoy with a continuous index i . The notation is introduced with reference to the observation of the convoy stability as its measurement of the quotient $V_K = \hat{\ddot{x}}_{i+1}(\omega)/\hat{\ddot{x}}_i(\omega)$ of the (complex) acceleration amplitudes. The convoy is stable precisely when the condition

$$|V_K| = |\hat{\ddot{x}}_{i+1}(\omega)/\hat{\ddot{x}}_i(\omega)| \leq 1, \quad \text{für } \forall \omega \geq 0 \quad (24.12)$$

is fulfilled. Otherwise, from a disturbance which is quite small, the frequency components of the frequencies for which this condition is not fulfilled will be greater with each subsequent convoy position. Convoy stability obviously applies for the idealized control principle represented in (32.11) because

$$|V_K| = |e^{-j\omega\tau_{\text{set}}}| = 1 \quad (24.13)$$

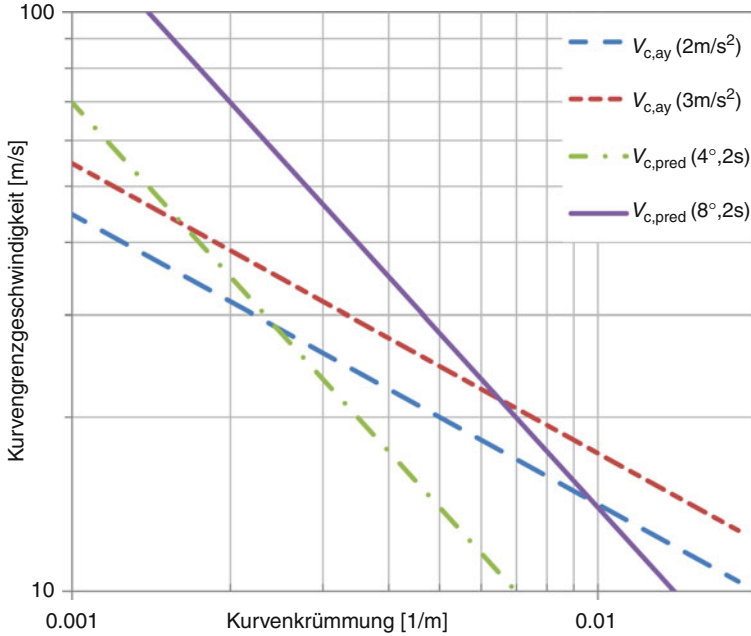
even if semi-stable without damping. This approach is not suitable in practice, but it illustrates a basic controller design. The disadvantages of this approach are the numerically unsuitable detection of the acceleration of the preceding vehicle (differentiation of the relative speed and the driver's actual travel speed, the required filtering leads to phase delay) and the fact that there is no correction opportunity if the speeds do not match or in the case of deviations in the distance.

For this purpose, the following is a control design based on relative speed:

$$\ddot{x}_{i+1}(t) = (\dot{x}_i(t) - \dot{x}_{i+1}(t))/\tau_v = v_{\text{rel}}/\tau_v \quad (24.14)$$

or in the frequency range

$$\hat{\ddot{x}}_{i+1}(s) = \hat{\dot{x}}_i(s)/(1 + j\omega\tau_v) \quad (24.15)$$



■ Fig. 24.15

Curve limit speeds for curve control depending on the curvature κ ($v_{c, ay}$ limit speed resulting from the limit lateral acceleration $a_{y \max}$, derived $v_{c, p}$ from the maximum azimuth angle ϕ_{\max} and the width of the look-ahead region)

With few steps, this approach can be transferred to an acceleration-led approach such as (24.11), wherein the acceleration value of the preceding vehicle is not delayed by a fixed time but is filtered in a PT1 element and, thereby, decelerated implicitly by τ_v . The application of (24.14) and/or (24.15) is obviously convoy-stable, but it only meets the requirements of a constant time gap if τ_v equals τ_{set} . Moreover, this control approach is not suitable for reducing any distance deviations. For this purpose, the controller is extended by an additive correction component for the relative speed which is proportional to the difference between the set and actual distances:

$$\ddot{x}_{i+1}(t) = \left(v_{\text{rel}} - \frac{d_{\text{set}} - d}{\tau_d} \right) / \tau_v \quad (24.16)$$

or in the frequency range

$$\hat{\ddot{x}}_{i+1}(s) = \hat{\ddot{x}}_{\ell}(s) \frac{1 + j\omega\tau_d}{1 + j\omega(\tau_d + \tau_{\text{set}}) - \omega^2\tau_d\tau_v}. \quad (24.17)$$

The stability condition $|V_K| \leq 1$ is now only met if τ_v is small enough:

$$\tau_v \leq \tau_{\text{set}}(1 + \tau_{\text{set}}/2\tau_d) \quad (24.18)$$

So far, the choice of the distance control time constant τ_d has been left open. For this purpose, a reference scenario can be used, namely falling back in a cut-in situation. In this case, it is assumed that the cutting-in vehicle cuts in without a speed difference at a distance which is 20 m smaller than the set distance. An appropriate reaction would be a delay of around 1 m/s^2 which equates to taking ones foot off the accelerator pedal or very slight braking. For such a response, according to (24.16), the product must be $\tau_v \cdot \tau_d = 20\text{s}^2$.

It follows from (24.16) that the smaller the vehicle-following time gap, the higher must be the loop gain defined by τ_v^{-1} for the relative speed. However, a high loop gain also means a minimum damping of speed fluctuations of the target vehicle for frequencies above 0.05 Hz.

8 Target-Loss Strategies and Curve Control

When negotiating bends, target loss is possible because the maximum azimuth angle of the ACC sensor is inadequate (see 5.2) to detect the target object. Even when traveling straight ahead, short-term target loss is possible if, e.g., reflectivity is low (for example, motorbike) or object differentiation fails. In these cases, immediate acceleration to a set speed, as after the cutting-out of the target vehicle, would be inappropriate. A differentiation is often made between these two scenarios in that in the cutting-out case, the target plausibility (see 6.3) deteriorates due to a negative assignment measurement ($LP < 0$) to the driving corridor and the object is still detected in the case of this “loss of target.” Conversely, when driving round narrow bends or in the case of other target losses, the response to which should not be rapid acceleration, the target loss is related to object detection errors and a positive assignment ($LP > 0$) to the driving path in the last known measurement. The response design differs when this differentiation is made: In the first case, acceleration after loss of target is brisk unless a new target object limits acceleration, while, in the second case, acceleration is initially suppressed. Yet, how long should this continue and what strategies then follow? The time gap preceding the loss of target is used for the duration of the suppressed acceleration. Should the target object disappear from the measurement range because it has entered a bend, this can be verified by the ACC vehicle after traveling the distance corresponding to the time gap because the curvature will be different from that at the time of the target loss. If this curve criterion is met, the acceleration suppression strategy can be replaced by a curve control. In the other case, it is assumed that the target object is no longer in the driving corridor, and the speed is then adjusted to the new situation.

In the case of curve control, two aspects are important: Lateral acceleration and the effective range $d_{\text{max,eff}}$ of the ACC sensor. This is given by the curvature κ and the maximum azimuth angle ϕ_{max} and equals approximately:

$$d_{\text{max,eff}} = 2\phi_{\text{max}}/\kappa \quad (24.19)$$

From this, a speed $v_{c,p} = d_{\max,\text{eff}}/\tau_{\text{preview}}$ can be deduced for the minimum time gap τ_{preview} available for an approach:

$$v_{c,p}(\kappa, \phi_{\max}, \tau_{\text{preview}}) = 2\phi_{\max}/(\kappa \cdot \tau_{\text{preview}}) \quad (24.20)$$

This speed enables the decision to be made as to whether to accelerate further. This strategy results in an appropriate driving strategy precisely for very tight bends, e.g., cloverleaf junctions.

The second criterion, presumably used in all ACC systems, is lateral acceleration. As in the derivation of the curve classification (► Sect. 5.2), a lateral acceleration limit $a_{y\max}$ that defines the comfort range is assumed which is between 2 m/s^2 (at higher speeds) and 3 m/s^2 (at lower speeds). From this, in turn, a curve limit speed $v_{c,\text{AY}}$ can be derived:

$$v_{c,\text{ay}}(\kappa, a_{y\max}) = \sqrt{a_{y\max}/\kappa} \quad (24.21)$$

Both limit speeds are illustrated for two typical values respectively in ► Fig. 24.15. If the actual driving speed is above the reference speeds, a positive acceleration is either reduced or even reversed (negative) without falling into the range of significant decelerations of more than 1 m/s^2 .

In combination with the “target-loss acceleration suppression” described above, astoundingly good “blind flights” of a proven quality of 80% were achieved in a series of tests (Luh 2007), measuring whether the driver continued the run without intervening following target loss.

The reaction to curve-related target losses can be improved using information from a digital map, ideally with precise lane accurate positioning (not yet supplied in production vehicles). The curvature is then detected in advance and adjusted to the control strategy at motorway exits.

Another challenge for the ACC developer is when the target vehicle turns off the carriageway. The change in direction of the speed vector of the preceding vehicle results in a perceptible deceleration for the following vehicle. As the sensor measures only this, the ACC vehicle deceleration is disproportionate and must be reduced in a suitable manner.


8.1 Approach Strategies

The approach capability is defined as the maximum negative relative speed $-v_{\text{rel,appr}}$, which can be controlled by ACC with respect to a vehicle traveling at a constant speed before a critical distance $d_{\text{appr,min}}$ is exceeded. It depends on the distance $d_{\text{appr,0}}$ at the start of the deceleration, on the assumed constant maximum increase of the deceleration $\ddot{x}_{v,\min} = -\gamma_{\max}$ and on the maximum deceleration = minimum acceleration $\ddot{x}_{v,\min} = -D_{\max}$.

$$-v_{\text{rel,appr}} = \sqrt{2D_{\text{max}} \cdot \left(d_{\text{appr},0} - d_{\text{appr,min}} + \frac{D_{\text{max}}^3}{6\gamma_{\text{max}}^2} \right) - \frac{D_{\text{max}}^2}{2|\gamma_{\text{max}}|}} \quad (24.22)$$

$$d_{\text{appr},0} = d_{\text{appr,min}} - \frac{D_{\text{max}}^3}{6\gamma_{\text{max}}^2} + \left(-v_{\text{rel,appr}} + \frac{D_{\text{max}}^2}{2|\gamma_{\text{max}}|} \right)^2 / 2D_{\text{max}} \quad (24.23)$$

The distance required for a noncritical approach increases approximately fourfold with the differential speed and approximately reciprocally with respect to the maximum deceleration. At a distance of 100 m at $D_{\text{max}} = 2.5 \text{ m/s}^2$ around 20 m/s (72 km/h), differential speed can be compensated for, while for an approach capability of $-v_{\text{rel,appr}} = 100 \text{ km/h}$, $d_{\text{appr},0} \approx 120 \text{ m}$ and $D_{\text{max}} \approx 3.5 \text{ m/s}^2$ are required.

The ramp of the deceleration reduces the approach capability but increases transparency for the driver, see also  Sect. 2.2.

Naturally, in the case of dynamic approaches, it is impossible to avoid values falling below the stationary reference distance and/or the reference time gap. Therefore, a significantly smaller reserve distance value can also be used as the reference distance $d_{\text{appr,min}}$ for a successful approach. It should be noted however that undershooting is only permissible over a distance of 250–300 m.

8.2 Reaction to Static Objects

Static objects may be obstacles lying in the driving corridor. However, most are irrelevant targets such as drain covers, bridges, or signs. Even at speeds of 70 km/h, deceleration of around 2.5 m/s^2 should be started some 100 m before the object. However, since the probability of error in the target selection is still very high in this case, a reaction to static objects is recommended only in exceptional cases. The most important exception relates to the history of static objects. If these are measured in advance with an absolute speed which can be distinguished from zero, these are classified as “halted” objects and may also be treated as potential target objects. Otherwise, the conditions for a response to static objects are limited to an immediate area of up to approximately 50 m. The response may be either suppression of acceleration, in which a speed increase is suppressed as long as the static object is detected in the driving corridor, or a warning that the vehicle will drive over the object.

9 Longitudinal Control and Actuators

9.1 Basic Structure and Coordination of Actuators

Longitudinal control presents the challenge of converting adaptive speed control, namely the ultimate reference acceleration obtained from various individual controllers, into an actual acceleration. For this purpose, the sum forces (or sum torques) of respectively self-

contained subordinate control circuits of the drive and brake survey systems are adjusted so that the desired acceleration can be implemented. Even though simultaneous actuation of a drive torque and a brake torque is possible, it is generally avoided, and the respective elements are controlled independently.

With respect to the design of harmonic transitions between drive and brakes, it makes sense to choose a physical value with which both actuator subsystems can be controlled equally. Wheel torques (but also wheel forces) are an option. In this case, a summary observation is sufficient, i.e., the sum of the wheel torques acting on all four wheels since ACC does not apply torques to individual wheels. In this way, coordination on the basis of the same physical signal can be effected as close to the actuators as possible, as shown in the following sections.

ACC requires the implemented actual sum wheel torque (and/or force) to calculate the driving characteristic equation and therefore, amongst other things, carry out a gradient estimate. The drive must also provide the current set maximum and minimum value as the sum wheel torque. In this case, particularly the minimum possible torque, i.e., the achievable torque in the current gear shift operation is important since the brake can only be activated when it is no longer possible to decelerate via the drive.

As outlined in more detail in later chapters, ACC control does not require absolute accuracy of the actuators since deviations from the required reference value can generally be well compensated for by the enclosed control loop. Only at the start of the control process and at transitions between the different actuators does good absolute accuracy simplify the control. In principle, however, good relative accuracy is necessary to achieve the desired control comfort.

9.2 Brake

The ACC systems without braking system intervention, which were initially supplied mainly by Japanese car manufacturers, were not well received in Europe because of the minimum deceleration caused solely by the engine drag torque in conjunction with gear shifting meaning that the driver had to actuate the brake too often. The pioneering equipping of top-class vehicles since 1995 with ESP and Brake Assist, which support the driver in emergency braking situations, has greatly simplified the implementation of a suitable brake intervention for ACC systems and hardly any ACC is now found without brake system intervention.

As an engine torque interface for the reference torque requirement had already been introduced for ASR and ESP systems, it made sense to use the specification of a brake torque as a brake parameter. This has the advantage that the distribution between the various actuators in the ACC controller is very simple and the specifics of the individual actuators have little or no influence on the controller design, which greatly simplifies transferability to different vehicles and models.

If one observes the transfer function of the brake, one recognizes that pressure changes with the factor 0.1 effect the deceleration, i.e., pressure increases of 1 bar are only just

below the detection threshold of 0.15 m/s^2 . Correspondingly high are the metering requirements for the brake actuator.

The following applies for the relationship between deceleration D (= negative acceleration) and braking pressure p_{Br} :

$$\Delta D = \Delta p_{\text{Br}} \frac{A_K \mu_{\text{Br}} R_{\text{Br}}}{m_v R_{\text{dyn}}} \quad (24.24)$$

Key to symbols:

ΔD Change in vehicle deceleration

Δp_{Br} Change of braking pressure

A_K Total area of the brake pistons

μ_{Br} Sliding friction coefficient between brake pad and brake disk

R_{Br} Effective radius of the brake disks (average value)

m_v Vehicle mass

R_{dyn} Dynamic radius of the wheels

$A_K \cdot \mu_{\text{Br}} \cdot R_{\text{Br}} = 70 \text{ Nm/bar}$ and $m_v = 2,100 \text{ kg}$ and $R_{\text{dyn}} = 0.34 \text{ m}$ can serve as approximate values. Transmission values of $\Delta D / \Delta p_{\text{Br}}$ $0.07 \dots 0.14 \text{ m/s}^2 \text{ bar}$ therefore result depending on the vehicle configuration.

9.2.1 Actuator Dynamics

For comfort functions such as ACC, deceleration variations of up to 5 m/s^3 are typically permitted (see [Sect. 2.2](#)). This would result in a required pressure variation dynamic of $30\text{--}40 \text{ bar/s}$. In order to be able to follow the specified torque and/or pressure trends sufficiently dynamically, the brake system must be able to follow changes of up to 150 bar/s .

Dynamic following of the reference value with sufficiently rapid pressure increase up to the start of braking and as far as possible delay-free following in the case of pressure modulations is required. The maximum time delays in this case should remain $< 300 \text{ ms}$. The preconditions for this, along with a correspondingly sized pump, are predominantly the choking of the hydraulic system in the pump intake area in order to be able to provide the required volume, largely irrespective of temperature. The control of the reference value must be free from disturbances as this will be perceived by the driver as extremely unpleasant. Together with rapid following of dynamic reference values, as far as possible, stepless following of small or slowly changing reference values is absolutely essential because precisely this kind of control of small control differences is typical for ACC operation. Stationary deviations are also to be avoided as these turn into speed and distance errors and can lead to limit-cycle oscillations.

9.2.2 Control Comfort

As already described in the introduction, the vehicle responds very sensitively to pressure changes. To ensure that a sensitive driver perceives the pressure increase as continuous, the

brake system must be able to handle braking increments of less than 0.5 bar. As far as possible, braking pressure increase and decrease should be silent, harmonious, and continuous. Inadvertent pressure changes of more than 1 bar are to be avoided. Additional pump elements are beneficial for an even pressure increase while continuously regulating valves are beneficial for pressure decrease. In terms of acoustics, a low pump speed is desirable, as are suitable location of the hydraulic unit and the suitable placement of brake leads in order to prevent the absorption of vibrations from the chassis. A complicating factor is that in the case of brake intervention, one of the substantial noise emitters in the vehicle, the engine, is reduced to its acoustic minimum, the drag range.

9.2.3 Feedback Information

The brake subsystem is the key supplier of internal vehicle state values, the most important of which are: vehicle speed, yaw rate, steering angle, brake light switch, slip control information. Also, the current actual brake moment is fed back in order that the ACC can carry out a gradient estimate. The ESP system provides binary state information (flags) (e.g., ABS-active, ASR-active (= Traction Control active), ESP-active) for an appropriate response to control states.

9.3 Drive

A review will now be made of the combination of the combustion engine and the automatic transmission, with the manual transmission treated as a special case. Combinations with hybrids are also conceivable. In principle, we can say that transitions between electric and combustion engines must be imperceptible as far as possible for the ACC as for the driver. The drive is, furthermore, only a torque provider for the ACC since how the torque is generated is irrelevant for the system function. With respect to recuperated braking with an electric machine, it is important to ensure corresponding coordination with the brake system, which has to implement the change-over to the friction brake.

It has proven useful to view the engine and transmission as one unit from the perspective of the ACC and to specify direct total wheel reference torques and to leave it to the drive system to decide how this torque is to be applied, either by changing the engine torque or by changing the gear ratio.

Consequently, a change of acceleration as observed for braking Δa is proportional to the sum wheel force change and/or the sum wheel torque change:

$$\Delta a = \frac{\Delta F_{R\Sigma}}{m_v} = \frac{\Delta M_{R\Sigma}}{m_v R_{dyn}} \quad (24.25)$$

Key to symbols:

Δa Change of vehicle acceleration

m_v Vehicle mass

R_{dyn} Wheel radius

While direct actuation of the motor via engine torque reference value is possible, specific action is required to influence the transmission in order to maintain an adequate dynamic while avoiding unwanted gear shifts. Simply using a heavily gearshift-based characteristic as in CC is inadequate because the ACC following controller must be far more dynamic in design than a vehicle speed controller purely designed for constant travel.

Equally, a direct conversion of the engine torque reference values into virtual driving pedal angles in order to control the transmission logic is not suitable because the ACC attempts to emulate a preset acceleration precisely and, other than with the driver, deviations are directly reflected in reference value changes which at certain operating points could lead to oscillating gear shifts.

9.3.1 Engine Control (Control Range, Actuator Dynamics, Steps/Accuracy, Feedback Information [Loss Torque of Ancillary Units])

As with the brake, the ACC requires access to the entire possible torque range for the necessary control range in order to cover all relevant driving situations. The required servo dynamics corresponds to the dynamics required by the driver, which should not present a problem in most modern systems because the driver reference values can also be transferred electronically; driver and ACC presettings are therefore principally transferred via the same path.

The drive optimally applies the required sum wheel torque of the ACC function (similar to the accelerator pedal) at the respective operating point. The engine, transmission and ancillary units are used to implement the reference value. Coordination is autonomous in the drive system as far as possible. If this is not supported, conversion to the engine torque by the ACC control unit or a longitudinal dynamic module is required, for which the current gear transmission must be known.

The ACC function differentiates between different operating modes for comfort purposes which relate to the coordination of the drive's different actuation options (e.g., overrun fuel cut-off, gearshifts, incorporation of ancillary units). Thereby, minor inconstancies in the torque application such as occur, for example, when activating the overrun fuel cut-off, are avoided or permitted. In addition, more serious inconstancies in the torque application such as occur, for example, during additional gear shifting in automated multistep reduction gears, are avoided or permitted.

Special Features of Combination with Manual Transmissions

The engine control determines the drive ratio of the wheel torque/crankshaft torque via the speed transmission of the different gears and thereby calculates an engine torque from

the drive requirement of the ACC function and implements this in the best possible manner.

During the shift process, i.e., after activation of the clutch by the driver, the engine control adjusts the crankshaft speed for synchronization of the crankshaft/transmission input speed in the target gear. The crankshaft speed reference value is determined depending upon the target gear predicated in the engine control unit.

The engine control evaluates the crankshaft speed and advises the driver, taking into account the situation, to select a lower gear. The request to select a higher gear is not necessary.

To prevent the engine from stalling, the engine must have the opportunity to switch off the ACC function if the driver does not respond to the shift instruction. ACC is also switched off if the clutch process exceeds a time limit (e.g., <10 s) or a suitable gear has not been engaged.

9.3.2 Transmission Control

The ACC state control requires, as one of its activation criteria, information from the transmission that a valid (forward) gear is engaged.

If engine torques are to be preset, the ACC needs the current torque amplification V_S from the transmission; this is the ratio of the force $F_{R\Sigma}$ on the drive axle to the engine torque M_M and is given by the product of converter amplification μ_W , the gear ratio i_G of the current gear, and the axle gear transmission i_A divided by the dynamic wheel radius R_{dyn} :

$$V_S = F_{R\Sigma}/M_M = \mu_W \cdot i_G \cdot i_A / R_{dyn} \quad (24.26)$$

In this case, the converter amplification is generally incorporated as a curve which may need to be temperature-compensated.

FSRA may also use electronic gear shift systems as additional protection for stopping management. In this case, the parking lock is engaged on detection of the intent to leave the vehicle. This is sufficient in combination with a multistage driver early-warning system which advises the driver of the responsibility to make the vehicle safe when it is stopped.

It is not sufficient, however, as the sole safety means for a fully automated stopping management system (without driver intervention) since the parking lock blocks only the propeller shaft. At corresponding μ -split conditions, the wheels could turn and the vehicle could roll away. Also in the case of a late request or in the case of an error when the vehicle is already rolling, it is not possible to safely engage the parking lock above approximately 3 km/h, while an EPB can function at any speed in principle.

10 Use and Safety Philosophy

10.1 Transparency of the Function

Transparency of the system responses is essential for acceptance of the ACC system. Only when users are quickly able to predict the system responses will they also use the system effectively. This presents the developer with the problem of making the control as simple as possible and sometimes leaving out features which an experienced user and of course the developer would value. As the driver gives up a part of the vehicle management task to the system when the ACC function is active, and needs only to monitor this, the transparency of the system plays an important role. Because current ACC systems only perform a part of the longitudinal control, it is recommended and in fact necessary to select system limits for normal use of the system so that they are reached and/or exceeded with a certain regularity. This ensures that the driver knows the system limits at any one time and is versed in reassuming control from the system.

Adaptive Cruise Control is not a safety function. It is designed first and foremost to increase driver comfort. Of course, a comfort system should not pose any dangers, therefore the ACC system must guarantee a level of safety corresponding to this requirement. Fault tree analyses have shown that dangerous situations can only occur if the driver does not employ intervention options. Two consequences then ensue:

1. The driver must not find resuming of control too demanding. In particular, the driver must recognize the need to resume control and take the corresponding steps and select the correct operating method in sufficient time.
2. The driver-control-take-over option must also be fault-tolerant so that this option, e.g., switching off the control, strong deceleration, or strong acceleration, can only be blocked in an extremely improbable manner.

Prompt recognition of the need to take over the controls is derived from the driver's perception gained from past experience. In particular, too much confidence in the technology because only error-free functioning is experienced would be problematic because the driver would be unprepared both for the occurrence and also the response. In the case of ACC, this difficulty is not an issue because, as explained above, it is impossible to perfect its function. This negative aspect therefore has the benefit that the driver is permanently trained for the error situation. The driver remains aware that she/he may have to resume control in the case of unwanted performance and is well practiced in when and how this is carried out.

10.2 System Limits

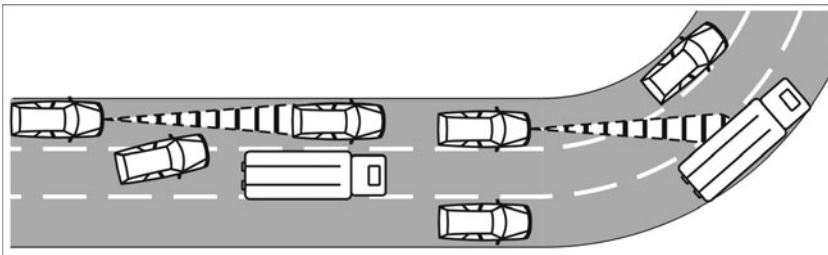
Beam sensors such as radar or lidar sensors offer precise detection of distance and relative speed and at least the radar sensors are largely robust to weather influences. Also, due to the limited opening angle and the difficulties involved in the lane assignment of the

detected object, especially in curve situations, there are limitations that sometimes lead to unexpected or incomprehensible system responses which must be explained to the user via suitable media.

Because of the narrow detection range of the ACC sensors, drivers cutting in directly in front of the actual vehicle are recognized late (► Fig. 24.16, left). The assignment of the detected objects is still problematic when entering a bend, especially when curve-travel cannot yet be detected because of the imminent vehicle signals (steering wheel angle, yaw rate) (► Fig. 24.16, right).

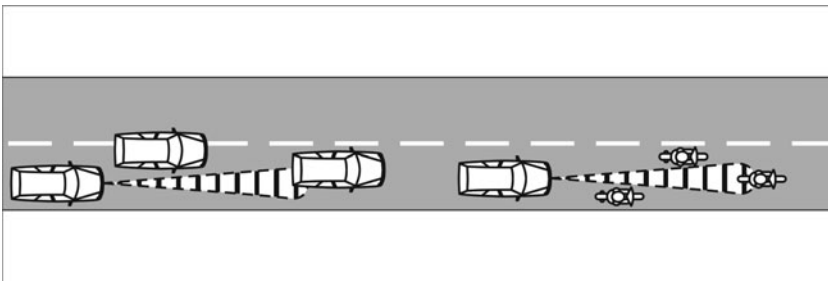
We can expect help from the use of cameras which are able to detect lane markings and by using information from modern navigation systems about the expected route of the road. Extremely off-center driving methods can also lead to failures in detection. Particularly in the case of motorcycles, detection is a problem due to their narrow silhouette (► Fig. 24.17).

Some of the weaknesses above relate to first-generation ACC systems and have been at least partially compensated for by the wider focal range of sensors of the successive generation or by the use of additional sensors with a smaller reach but a significant lateral detection range, as are increasingly used in FSRA systems.



■ Fig. 24.16

Typical problem situations; *left*: late reaction to driver cutting in; *right*: difficulty in assigning objects when entering a bend



■ Fig. 24.17

Typical problem situations: Ambiguity in the case of off-center motor vehicles and motorcycles

11 Safety Concept

The failure tolerance of the driver-control-take-over option has been eased by the distribution of the system. One actual example is the reading of the brake pedal switch both by the engine control as well as by the ESP. On detection of pressure on the brake pedal, the torque request of the ACC longitudinal control is ignored by the engine control. Deceleration requests to the brake control are also suppressed when pressure is applied to the brake pedal. Both the actuation of the brake pedal and the accelerator pedal are detected as redundant, so that both the actuation as well as the subsequent response states for one-off errors are protected even if the ACC-ECU or the data network is interrupted.

Since the apportionment of the tasks can vary greatly, as the specific examples show, there are no real standard solutions. Instead, the guaranteed driver intervention option is demonstrated by a fault tree analysis.

Along with the permanent availability of the driver intervention options, inherent safety of the ACC system is essential. Again, the deployment of the system proves to be an advantage. For example, the ESP system as an evidently inherently safe system can be used to monitor the ACC control. If one selects the resulting vehicle acceleration as the monitoring value, all theoretically possible fault sources are included. Since first-generation ACC systems have very narrow limits, generally $+1 \text{ m/s}^2$ and/or -2 m/s^2 , this kind of acceleration monitoring is easy to implement. The only disadvantage is that in this kind of monitoring, the acceleration and/or deceleration is applied to the vehicle for a short time before it is suppressed by the ESB. However, the limits can be selected to ensure that over 95% of normal drivers are able to cope with this.

12 Users and Acceptance Studies

12.1 Acceptance

The results of the testers in all studies of acceptance carried out so far are unambiguous.

Becker and Sonntag (1993) describe how the testers in the pilot study perceive driving with ACC subjectively as safer, more relaxing, and less stressful than manual driving. This conviction was arrived at despite the prototype status of the test vehicles, which exhibit certain serious sensor flaws. Nevertheless, the expectations of the system of the trial participants were fully met and at times exceeded. It is clear, therefore, that the results of the testers with respect to acceptance and comfort at this level of maturity of ACC are largely robust.

Even with ACC systems without brake intervention, testers in the UMTRI study expressed themselves highly satisfied, which (Fancher et al. 1998) attribute to the reduction of “throttle-stress.”

By investigating the quality of processing, ancillary tasks (Nirschl and Kopf 1997) detect less mental stress on the driver when using ACC. In subjective statements, high acceptance is expressed and ACC is seen more as a comfort than a safety system.

As well as global satisfaction and acceptance on the part of drivers, (Weinberger 2001) analyses the temporal trends in long-distance journeys. All aspects such as “enjoyment of the system,” “intuitiveness in use,” “confidence in the controls,” “feeling of well-being,” and “stress” are rated in principle as good to very good. Over the duration of the trial, initial euphoria is replaced by a phase of relative disenchantment which nevertheless results at the end of the test in a significantly better evaluation than at the start.

12.2 Use

The object of any investigation is the time gap behavior of drivers in comparisons between manual driving and driving with ACC. In straightforward following situations (Abendroth 2001), the average minimum time gap is 1.1 s with both manual driving and ACC. In contrast, Becker and Sonntag (1993) find a proliferation of time gaps of 1.7 s with manual drivers, albeit with significant scatter of results. A possible explanation to this may be the more winding test route. In ACC mode, the average time gap is 1.5 s, which was specified in the pilot study as a basic setting. Filzek (2002) finds that when given free choice of levels 1.1, 1.5, and 1.9 s, testers choose an average ACC time gap of 1.4.

A significantly shorter average time gap of 0.8 s in manual driving is reported by Fancher et al. (1998). This apparent contradiction reveals the difficulty of transference between studies carried out in different traffic systems, in this case, USA and Germany.

Significantly, in all studies, polarization takes place with respect to the selected ACC time gap. While at the start, the testers “play” with the levels, the frequency of adjustment decreases over the duration of the test. Respectively, around half of the testers then choose either slightly lower or slightly higher levels. With respect to the frequently selected short time gaps, a limit to at least 1.0 s appears to make sense on safety grounds.

Fancher et al. (1998) investigated the time gap behavior in more detail and found that similar levels of 1.1, 1.5, or 2.1 s were selected depending on the age of the tester, i.e., older drivers selected correspondingly larger ACC time gaps.

Both Filzek (2002) and Fancher et al. (1998) report that significantly fewer drivers choose very small time gaps of below 0.6 s with ACC (Fancher et al.: 6 out of 108 testers).

Mercedes-Benz market researchers surveyed customers in the USA about the use of Distronic, see ► Fig. 24.18. The details relate to the S-class (W220, 1998–2005) and the SL (R230, since 2001). The rate of use in multiple-lane highways is expected to be substantially higher than in other road categories. The discrepancies between sports cars and saloons are surprisingly low in terms of the rate of use. Differences in types of use are somewhat greater. Since in the case of the Distronic operating concept the time gap stays at the old value purely mechanically, changing the time gap is only necessary if there is a reason for a change. Little or no use is made of this option. The distance setting is usually specified as average.

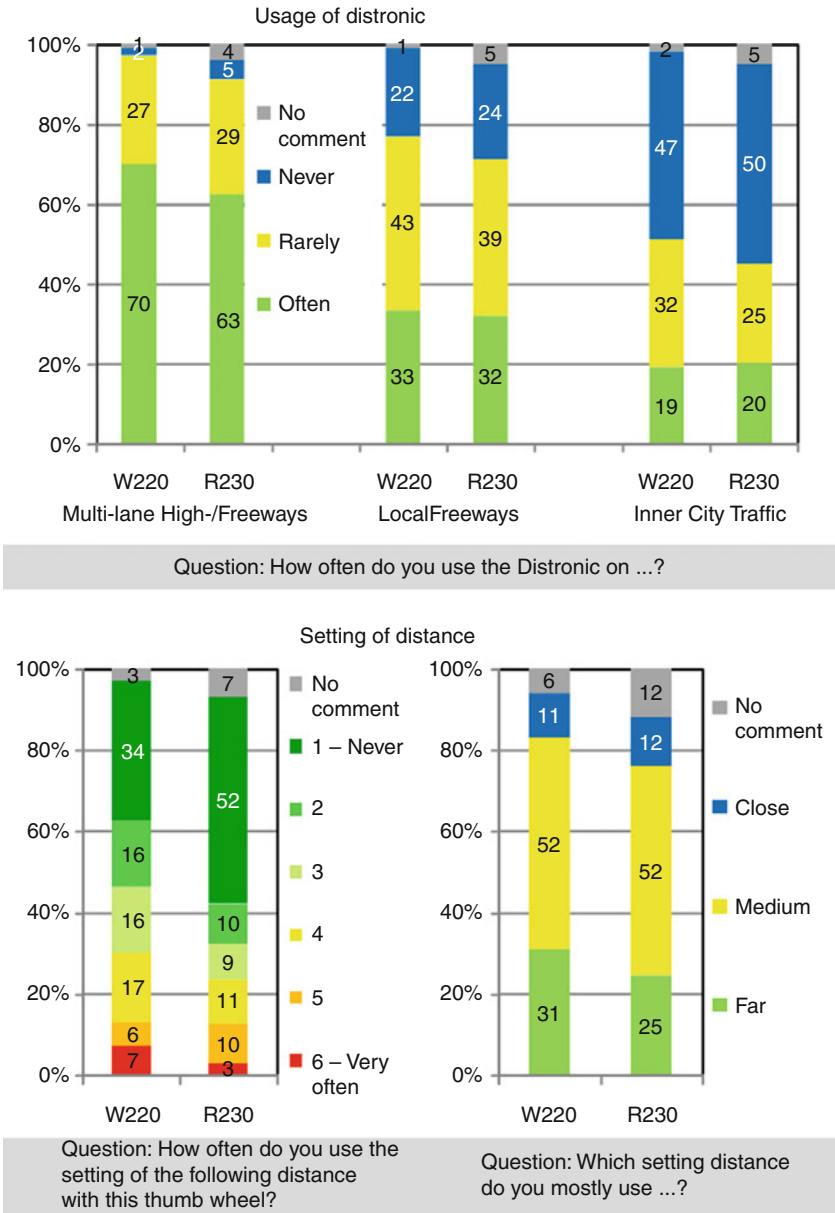


Fig. 24.18
Information about the use of an ACC system in the USA based on the DISTRONIC (Source: Mercedes-Benz Market Research 2005)

12.3 Driver-Control-Take-Over Situations

The principle simplicity of the driver’s mental perception of ACC according to Becker et al. (1994) is also attributable to the fact that a correct response to system limits in

control-take-over situations is possible even after using the system for a very short time. Fancher et al. (1998) explain that 60% of testers felt able to recognize control-take-over situations promptly and correctly after just 1 day. Ninety-five percent of testers agreed to this statement after 1 week.

Nirschl et al. (1999) also report that most testers were able after a short time to estimate which ACC situations required intervention. However, the average of the three analyzed ACC variants with a somewhat smaller brake delay of 1 m/s^2 lead to greater uncertainty in the estimation than the variants with stronger braking intervention and/or with no braking intervention.

Weinberger (2001) explains that the estimation of control-take-over situations was perceived subjectively by testers as noncritical and was perceived as simpler by the drivers as the period of use extended. The testers also stated that decisions are particularly easy in those situations which ACC cannot manage in principle (e.g., braking behind a halted vehicle). It was shown that after the driver takes over the controls, in 80% of cases the mean deceleration of the vehicle is below or equal 2 m/s^2 . This period is also covered by ACC so that one may conclude that these situations are also noncritical from an objective standpoint.

With a few exceptions, the first ACC vehicle tests show a uniform trend, although there would be sufficient reason to justify differentiation in the results:

- The technology of the systems analyzed differed considerably both in terms of the scope of functions and level of maturity.
- The traffic conditions in the USA are only partly comparable with those in Europe.
- Short-term and long-term tests were carried out, and in the long-term tests, clear learning effects were found which invalidate the predictions of some of the results of short-term tests.

Obviously, at least in its basic functions, ACC appears to be robust with respect to the stated differences in the performance of the tests. The core function was understood by the drivers at the start, irrespectively of the limitations of the predecessor systems.

With respect to the FSRA control-take-over situation, Neukum et al. (2008) analyzed a crossroads problem situation. A vehicle which had been at a standstill for a long period at a crossroads is initially concealed by the target vehicle and then overtaken shortly before the approach so that it is suddenly in the driving corridor of the FSRA vehicle. As a static vehicle which has not yet moved into the focal range of the radar, this is not accepted by the FSRA system as a control object, i.e., the driver must intervene and brake in order to avoid a collision. All testers were able to do so without the codrivers (present for this purpose) having to initiate the action. Nevertheless, this situation is rated by many drivers as threatening when encountered for the first time.

12.4 Comfort Assessment

The focus of the analyses documented in Didier (2006) was the investigation of comfort. For this purpose, two vehicles from different manufacturers with different ACC systems were driven by a total of 36 test persons. The subjective assessments, which were obtained by means of a questionnaire on selected comfort criteria, were compared. Although the systems in both vehicle series carried out the same functions, it was possible to detect even slight differences between the two systems with respect to comfort. However, the simultaneous analyses of the objective quantitative characteristics “Frequency of driver override by actuation of accelerator” and “Interruption of the control by driver brake intervention” could not be adequately equated with the comfort evaluation.

13 Conclusion and Outlook

13.1 Current Developments

With the series introduction of FSRA, everyday traffic situations are covered by the function range. The trend is toward providing the existing systems after the (normally top down) launch in smaller and less expensive models. Alternatives to the formally predominating radar sensor principle may be used, e.g., lidar systems. The current developments in environment sensors also indicate the potential of radar, and it is unlikely that this technology will be abandoned. Since the target vehicle segment is very price-sensitive, multi-sensor solutions (e.g., for FSRA) are currently reserved for vehicles in the top price range. Also the restrictive permission policy with respect to 24 GHz-UWB radar sensors greatly hinders their wider distribution.

In the course of further cost reductions, investigations are being carried out with simpler sensor arrays (e.g., 24 GHz mid-range sensors). Nevertheless, the reduced system costs mean function restrictions which so far, in comparison to the current standard ACC, have been serious enough to render series production out of the question.

Work is being carried out on concepts as illustrated in Winner and Hakuli (2006) or Mayser and Steinle (2007) which integrate the driver directly into the overall longitudinal guidance function.

13.2 Function Enhancements

Future interest lies predominantly in coupling active (radar/lidar) sensors with camera-based sensors. Cameras to detect the vehicle and environment situation have many potential uses: From automatic light control through road sign detection and lane marking and lane location detection to object recognition and classification. Fusion

with the data from additional sensors will provide a reliable situation assessment which will permit increased functions, e.g., vehicle following in traffic queues with automatic longitudinal and lateral guidance. It will also improve the quality of collision warning and/or mitigation system functions, see Steinle et al. (2006).

Many types of applications become possible with the aforementioned 24 GHz-UWB radar sensors: from parking assist to parking guidance systems, blind-spot monitoring, and even radar “all-round protection.”

Acknowledgments

This chapter is an extract from Winner H, Danner B, Steinle J (2009) Adaptive Cruise Control. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme (Handbook of Driver Assistance Systems), Chap. 32, Teubner Vieweg.

References

- Abendroth B (2001) Gestaltungspotentiale für ein PKW-Abstandsregelsystem unter Berücksichtigung verschiedener Fahrertypen, Dissertation TU Darmstadt, Schriftenreihe Ergonomie, Ergonomia-Verlag, Stuttgart
- Ackermann F (1980) Abstandsregelung mit Radar. Spektrum der Wissenschaft Juni 1980:24–34
- Becker S, Sonntag J (1993) Autonomous Intelligent Cruise Control – Pilotstudie der Daimler-Benz und Opel Demonstratoren. In: TÜV Rheinland Prometheus CED 5, Köln
- Becker S, Sonntag J, Krause R (1994) Zur Auswirkung eines Intelligenten Tempomaten auf die mentale Belastung eines Fahrers, seine Sicherheitsüberzeugungen und (kompensatorischen) Verhaltensweisen. In: TÜV Rheinland Prometheus CED 5, Köln
- Didier M (2006) Ein Verfahren zur Messung des Komforts von Abstandsregelsystemen (ACC-Systemen), Dissertation TU Darmstadt, Schriftenreihe Ergonomie, Ergonomia-Verlag, Stuttgart
- Fancher P et al (1998) Intelligent Cruise Control Field Operational Test. University of Michigan Transportation Research Institute (UMTRI) Final Report, Michigan
- Filzek B (2002) Abstandsverhalten auf Autobahnen – Fahrer und ACC im Vergleich, Dissertation TU Darmstadt. In: VDI Fortschritt-Berichte Reihe 12, Nr. 536, VDI-Verlag, Düsseldorf
- Furui N, Miyakoshi H, Noda M, Miyauchi K (1998) Development of a Scanning Laser Radar for ACC. In: Society of Automotive Engineers SAE Paper No. 980615, Warrendale, Pennsylvania
- ISO TC204/WG14 (2002) ISO 15622 Transport information and control systems – Adaptive Cruise Control systems – Performance requirements and test procedures
- ISO TC204/WG14 (2008) ISO 22179 Intelligent transport systems – Full speed range adaptive cruise control (FSRA) systems – Performance requirements and test procedures
- Luh S (2007) Untersuchung des Einflusses des horizontalen Sichtbereichs eines ACC-Sensors auf die Systemperformance, Dissertation TU Darmstadt, VDI Fortschritt-Berichte Reihe 12, VDI-Verlag, Düsseldorf
- Mayser Ch, Steinle J (2007) Keeping the driver in the loop while using assistance systems. In: SAE World Congress 2007 SAE 2007-01-1318, Detroit, Michigan
- Meyer-Gramcko F (1990) Gehörsinn, Gleichgewichtssinn und andere Sinnesleistungen im Straßenverkehr. Verkehrsunfall und Fahrzeugtechnik 3:73–76
- Mitschke M, Wallentowitz H, Schwartz E (1991) Vermeiden querdynamisch kritischer

- Fahrzustände durch Fahrzustandsüberwachung. In: VDI Bericht 91. VDI-Verlag, Düsseldorf
- Neukum A, Lübbecke T, Krüger H-P, Mayser C, Steinle J (2008) ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In: Proceedings 5. Workshop Fahrerassistenzsysteme, Walting
- Nirschl G, Kopf M (1997) Untersuchung des Zusammenwirkens zwischen dem Fahrer und einem ACC-System in Grenzsituationen. In: Conference Der Mensch im Straßenverkehr, Berlin 1997, VDI Bericht 1317, VDI-FVT, VDI-Verlag, Düsseldorf
- Nirschl G, Blum E-J, Kopf M (1999) Untersuchungen zur Benutzbarkeit und Akzeptanz eines ACC-Fahrerassistenzsystems. In: Fraunhofer Institut für Informations- und Datenverarbeitung IITB Mitteilungen
- Pasenu T, Sauer T, Ebeling J (2007) Aktive Geschwindigkeitsregelung mit Stop&Go-Funktion im BMW 5er und 6er. In: ATZ 10/2007, Vieweg Verlag, Wiesbaden, pp 900–908
- Prestl W, Sauer T, Steinle J, Tschernoster O (2000) The BMW active cruise control ACC. In: SAE World Congress 2000 SAE 2000-01-0344, Detroit, Michigan
- Steinle J, Toelge T, Thissen S, Pfeiffer A, Brandstätter M (2005) Kultivierte Dynamik – Geschwindigkeitsregelung im neuen BMW 3er. In: ATZ/MTZ extra, Vieweg Verlag, Wiesbaden, pp 122–131
- Steinle J, Hohmann S, Kopf M, Brandstätter M, Pfeiffer A, Farid N (2006) Keeping the focus on the driver: the BMW approach to driver assistance and active safety systems that interact with vehicle dynamics. In: Proceedings of FISITA World Automotive Congress, FISITA F2006D185, Yokohama, Japan, 22–27 Oct 2006
- Watanabe T, Kishimoto N, Hayafune K, Yamada K, Maede N (1995) Development of an intelligent cruise control system. In: Proceedings of the 2nd ITS World Congress in Yokohama, Yokohama, Japan, pp 1229–1235
- Weinberger M (2001) Der Einfluss von Adaptive Cruise Control Systemen auf das Fahrverhalten, Dissertation TU München. In: Berichte aus der Ergonomie, Shaker-Verlag, Aachen
- Weinberger M, Winner H, Bubb H (2001) Adaptive cruise control field operational test – the learning phase. In: JSAE Review 22, Elsevier, Amsterdam, p 487
- Winner H (2003) Die lange Geschichte von ACC. In: Proceedings Workshop Fahrerassistenzsysteme, Leinsweiler
- Winner H (2005) Die Aufklärung des Rätsels der ACC-Tagesform und daraus abgeleitete Schlussfolgerungen für die Entwicklerpraxis. In: Proceedings Fahrerassistenzworkshop, Walting
- Winner H, Hakuli S (2006) Conduct-by-wire – following a new paradigm for driving into the future. In: Proceedings of FISITA World Automotive Congress, Yokohama, Japan, 22–27 Oct 2006
- Winner H, Luh S (2007) Fahrversuche zur Bewertung von ACC – Eine Zwischenbilanz. In: Bruder R, Winner H (eds) Darmstädter Kolloquium Mensch & Fahrzeug – Wie objektiv sind Fahrversuche? Ergonomia-Verlag, Stuttgart
- Winner H, Olbrich H (1998) Major design parameters of adaptive cruise control. In: AVEC'98 Paper 130, Nagoya, Japan
- Winner H et al. (2003) Fahrversuche mit Probanden zur Funktionsbewertung von aktuellen und zukünftigen Fahrerassistenzsystemen. In: Landau K, Winner H (eds) Fahrversuche mit Probanden – Nutzwert und Risiko, Darmstädter Kolloquium Mensch & Fahrzeug, 3–4 April 2003, TU Darmstadt, VDI Fortschritt-Berichte Reihe 12, Nr. 557, VDI-Verlag, Düsseldorf
- Witte S (1996) Simulationsuntersuchungen zum Einfluss von Fahrerverhalten und technischen Abstandsregelsystemen auf den Kolonnenverkehr. Dissertation Universität Karlsruhe, Karlsruhe, p 23

25 Forward Collision Warning and Avoidance

Markus Maurer

Technische Universität Braunschweig, Institut für
Regelungstechnik, Braunschweig, Germany

1	<i>Introduction</i>	659
1.1	Motivation and Early Research Approaches	659
1.2	Definitions and Abbreviations	661
2	<i>Machine Perception for FCW and FCA</i>	662
3	<i>Approaching a Definition of FCX-Systems</i>	663
4	<i>Design Parameters of Actual Realizations</i>	665
4.1	CU-Criterion	665
4.2	Fundamentals of Driver Warning	666
4.3	Levels of Assistance in Dangerous Situations	668
4.3.1	FCC-Systems	669
4.3.2	FCW-Systems	669
4.3.3	Rear Impact Countermeasures	670
4.3.4	FCM-Systems	670
4.3.5	FCA-Systems	671
5	<i>Levels of Assistance in an Actual Realization</i>	671
5.1	Survey of Different Realizations by the Car Manufacturers	672
6	<i>System Architecture</i>	672
6.1	Functional System Architecture	675
7	<i>Design Process</i>	676
7.1	Systematic Design of Driver Assistance Systems	676
7.2	Example: Systematic Design of an “Automatic Emergency Brake”	678
7.2.1	User-Oriented Definition of the Function	678
7.2.2	Functional Tests of Driver Assistance Systems	683

7.2.3

Test Case “Justified Intervention”: Vehicle-in-the-Loop

684

7.2.4

Error Probability for “Unjustified Interventions”: Trojan Horses

684

8

Conclusion

685

Abstract: Forward collisions represent a significant portion of all severe accidents. This is why appropriate warning and collision avoidance systems are of great importance to increase traffic safety. Different system specifications are subsumed under the so-called FCX-systems; they differ in their way of affecting the overall system driver–vehicle–environment as forward collision conditioning, forward collision mitigation, forward collision warning, and forward collision avoidance systems. A more specific definition of FCX-systems is derived by distinguishing them from other related systems as adaptive cruise control and pedestrian safety systems which can also have an impact on forward collisions. The specifications of the actual systems already on the market can only be understood if the characteristics of machine perception are considered carefully. The progress in the field of machine perception enables the forward collision warning and avoidance systems. There are still limitations of state-of-the-art perception systems compared to attentive human drivers which must be considered when designing FCX-functions. The state of the art in FCX-systems is sketched highlighting realized examples of FCX-systems of different car manufacturers. The last focus is on a systematic design process which is recommended for driver assistance systems. The motivation for the assistance is always to be derived from accident research. Already in the early conceptual phase functional safety, legal, ergonomic, and marketing aspects should be taken into consideration. Only if a consistent functional specification is found, further developments including package and architecture aspects are justified. Concepts for testing and evaluation should be designed in an early development phase as well.

1 Introduction

1.1 Motivation and Early Research Approaches

Forward collisions represent a significant portion of all severe accidents. This is why systems for obstacle and collision warning have been included in the recommendations of the eSafety Support, one of the leading European initiatives (eSafety 2010). They list systems with high efficiency regarding traffic safety and a high impact on the annual fatalities on the road (Gelau et al. 2009, S. 26).

Detailed analyses of accidents have revealed that many drivers either do not brake at all or do not make use of the full deceleration capacity of their braking systems. ● [Figure 25.1](#) shows the percentage of the drivers who either applied the brake just comfortably or even missed braking altogether although an emergency braking would have been the adequate reaction. Their portion in percent is plotted over the severity of the accidents quantified by the so-called Maximum of the Abbreviated Injury Scale (MAIS), which takes the most severe injury the driver suffered as a measure.

Early publications from the 1950s described prototypical systems designed to warn the driver of forward collisions. General Motors built up a prototype that perceived the relative velocity and the distance to the leading vehicle with an airborne radar system.

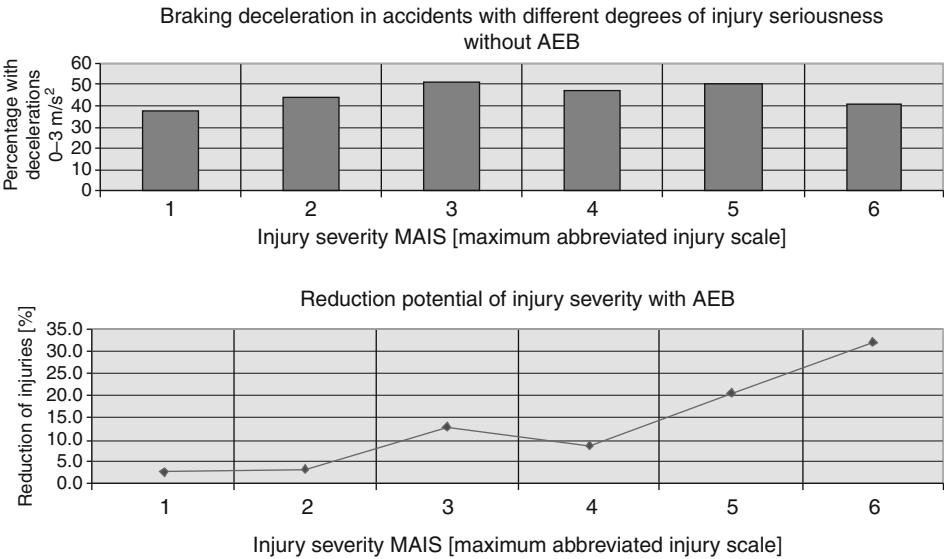


Fig. 25.1
Theoretical potential of an automatic emergency brake: decelerations at accidents with different degrees of severity of the injuries (nonassisted driving) (Kopischke S, 2000, personal communication, Ingolstadt); AEB automatic emergency brake, MAIS Maximum of the Abbreviated Injury Scale

Both state variables were indicated at the dashboard (Wiesbeck 2006). It took another 40 years to industrialize appropriate radar systems in such a way that they could be produced within economic constraints at least for small numbers of luxury cars.

In the 1970s, Bubb realized a system that displayed the braking distance necessary to stop safely in front of an obstacle depending on the current friction coefficient. In order to avoid any distraction of the driver, the braking distance was to be indicated in a contact analogous head-up display projecting the distance in the real scene in front of the vehicle (Bubb 1981). Again it took almost 40 years for this kind of head-up display to mature for the mass market.

Electronic controllable braking actuators were made available in a high percentage of new vehicles by the introduction of vehicle dynamics control systems in the 1990s. For the first time, supporting braking was feasible without equipping cars with additional actuators.

These actuators were exploited by the so-called hydraulic brake assist (HBA). Depending on the velocity with which the driver hits the brake pedal, the drive intention was to be determined. If an emergency situation was recognized, additional braking force was to be triggered by the HBA system (Kiesewetter et al. 1997). In practice, it turned out that the driver intention can be assessed only in rare occasions if just the drivers' pedal velocity is taken into account. Sporty drivers hit the brake pedals

so dynamically even in standard situations that they can hardly be distinguished from actions of average drivers in emergency situations. In order to avoid inadequate interventions of the HBA system, the triggering thresholds of these systems are adjusted conservatively nowadays. Therefore, average drivers can only be supported by HBA systems in selected emergency situations. Depending on the design principle, HBA cannot assist if the driver does not brake at all.

By introducing radar systems for adaptive cruise control, the technological base for machine perception of the outside world has been integrated into modern vehicles. Based on these sensors, systems were proposed in the 1990s which automatically perceive the environment in front of the vehicle and trigger an emergency braking once an accident cannot be avoided any longer within the limits of handling (Kopischke 2000).

1.2 Definitions and Abbreviations

Until today, a variety of different systems has been derived from the first approaches to obstacle and collision warning with machine perception. Each system was designed to protect the driver in special types of accidents. A few classifying concepts are defined below which separate the systems into different categories. In addition, a few abbreviations are introduced for easier reading.

Active safety: Collision avoidance is called active safety as well (definition after Naab and Reichart 1998).

FC: Forward collision.

FCA: Forward collision avoidance.

FCC: Forward collision conditioning.

FCM: Forward collision mitigation.

FCW: Forward collision warning.

FCX-systems: Systems which react in an appropriate way to reduce the impact of an impending forward collision on the passengers of a vehicle.

Forward collision: Collision in which the vehicle to be assisted crashes head-on into an obstacle or another road user.

Forward collision conditioning systems: Systems which condition subsystems in the vehicle in a way that, once triggered, they react faster (brakes), more effectively (brakes, seatbelt pretensioners) or more gently (smart airbags).

Forward collision mitigation systems: Systems which reduce the severity of an accident by applying appropriate countermeasures.

Forward collision warning systems: Systems which warn the driver of impending forward collisions.

Forward collision avoidance systems: Systems which avoid an impending forward collision by influencing vehicle dynamics.

Passive safety: Collision mitigation is called passive safety as well (definition after Naab and Reichart 1998).

2 Machine Perception for FCW and FCA

One reason for the emergence of the wide variety of current system collections is that in many aspects machine perception is inferior to the perception of attentive human drivers. It has been pointed out above that the systems for machine perception are the enabling technology for automotive driver assistance.

In the following, the characteristic features of machine perception systems are explained by a simple comparison: Driver assistance systems with machine perception are contrasted with the so-called conventional driver assistance systems which rely on direct measures or model based observations.

Conventional driver assistance systems support the driver in situations which are easy to measure or to estimate. Antilock braking systems intervene when a wheel is about to block. This can be determined by a conventional wheel speed sensor.

A vehicle dynamics control system brakes single wheels when the estimated sideslip angle exceeds an experimentally obtained threshold. In a sense, vehicles dynamics control systems already include tasks of machine perception: The estimation of the friction coefficient is a very challenging perception task especially if it is to be solved on time in order to adapt the current velocity of the vehicle to the road conditions.

A similar distinction between the two types of driver assistance systems is made in the code of practice for the so-called advanced driver assistance systems (ADAS): In contrast to conventional driver assistance systems, ADAS are equipped with sensors for the detection and interpretation of the environment of the vehicle (Donner et al. 2007).

To qualify as a driver assistance system with machine perception, support has to be given in situations *recognized* automatically. In an adaptive cruise control (ACC) system, radar reflections are interpreted as vehicles once these reflections fulfill certain temporal and spatial criteria. In a lane departure warning system, bright–dark transitions in the video image matching a specific Gestalt are understood as a lane with its markings. In a sense, machine perception means the ability to interpret automatically. In the current state of the art in this field, machine perception enables unprecedented possibilities of interpretation, but also unprecedented ways of misinterpretation.

When designing an innovative assistance system, the functional limitations of the state-of-the-art machine perception are to be taken into consideration from the very beginning. A successful strategy for the system design may put up with significant misinterpretations (e.g., one unjustified activation per 10,000 km in a safety system) if the system reaction is designed in a way that neither irritates nor endangers the driver. As an example, the design of an automatic warning jerk will be discussed later on (► Sect. 7.2).

As soon as an automatic intervention significantly influences the vehicle dynamics based upon machine perception, false automatic reactions are not accepted at all. In the automobile industry, there is currently no general agreement on how to define an appropriate false alarm rate given the impact of the machine intervention. The upcoming ISO norm 26262 aims at establishing the methods to develop innovative systems depending on a hazard analysis and a risk assessment.

In order to increase the reliability of the overall system, machine perception systems are equipped with redundant sensors whose data are fused to an internal environmental representation as consistently as possible. Interventions are only allowed in situations which can be specified formally so that erroneous automatic interpretations become very unlikely. In addition, the tracking of the traffic situation as it is developing over time is exploited in order to verify the machine interpretation. If in doubt, the assisting action is suppressed in order to avoid false reactions which could endanger any traffic participant. In the design of safety systems, this is also called a conservative systems design. This demand on redundancy is supported by a legal line of arguments which attempts to assess new systems by looking for analogies. Lawyers could argue that also in vehicle dynamics systems important state variables are perceived with redundant or at least functionally redundant sensor systems.

Given a special situation to be intervened in, the robustness of the machine perception can be further improved by tailoring both sensors and signal processing for this special task. Even though this principle is widely spread in nature where evolution supports the individuals best adapted to their ecological niche, a similar way of tailoring machine perception systems to special tasks is a major barrier for reusing sensors and systems in other assistance systems.

3 Approaching a Definition of FCX-Systems

A more specific definition of FCX-systems can be derived from distinguishing them from other related systems.

In contrast to ACC (🔗 Chap. 24). In principle, the comfort system ACC and FCX-systems are discussed as independent systems. However, they are coupled both technologically and even with respect to their impact to real-world accidents. As already shown in the introduction, the radar system of ACC serves as the technological base for the machine perception of FCX-systems in series production cars.

There is an ongoing controversial discussion how ACC affects traffic safety in general and forward collisions in particular. Users reported that, by intervening automatically, ACC had warned them of dangerous situations or avoided accidents directly.

Considering these single cases, ACC works as an FCX-system in a sense. When using ACC, many drivers accept longer distances to the leading vehicles compared to their normal style of driving.

A reduction of forward collisions is expected as long as the driver supervises the system attentively. There is still a need of further statistical analyses quantifying the impact of ACC as an FCX-system. There is a growing discussion among lawyers whether drivers should use ACC mandatorily if available (Vogt W, 2010, Bergisch Gladbach, personal communication).

Apart from the positive impact just sketched on traffic safety, system developers have made sure from the very beginning that the use of an ACC system would not affect the safety of the vehicle negatively. The fundamental principles of the concerns are derived

from earlier experiences with automating plants and aircrafts (Bainbridge 1983), or from basic research in psychology (Yerkes–Dodson Law). In simple terms, the evident experience is stated that the mental workload should not be further reduced when the driver is already bored. As long as the driver is responsible for the driving task, it has to be ensured that he is sufficiently involved in the vehicle guidance task.

Buld et al. (2005) demonstrated in a driving simulator that the drivers' performance related to their driving task would rather decrease than improve with the perfection of the ACC system due to progresses in the technical development (Buld et al. 2005, p. 184). When driving with ACC the driver may get tired faster than a nonassisted driver.

Therefore, it is a significant milestone in the development process of car manufacturers to test the usability of ACC intensively when developing any innovative system variant. If any doubt about the usability remains, the system variant will be modified to ensure that there are no negative impacts on traffic safety (Neukum et al. 2008). As another aspect, the usability of ACC during long-term operation was also analyzed and intensely documented for the first time by Weinberger (2001).

In contrast to proactive pedestrian protection systems. Formally, there is a big overlap of proactive pedestrian protection systems and FCX-systems; forward collisions with pedestrians are a very important type of accidents. The diversification of the systems is again caused by the limited possibilities of machine perception and the resulting highly specialized approaches to machine perception (► Sect. 2).

FCX-systems are designed first of all to protect the passengers of the assisted car. Therefore, the recognition of other vehicles even far ahead of the vehicle is of major importance. The proactive pedestrian protection systems primarily guard the pedestrians outside the assisted vehicle. Therefore, the specialized machine perception has to recognize explicitly pedestrians and react appropriately to their needs. In this meaning, proactive pedestrian protection systems are specialized FCX-systems. Even though there is no explicit representation of pedestrians, FCX-systems can also guard pedestrians by recognizing them as relevant objects (but not explicitly as pedestrians!) and reacting properly.

In contrast to integrated safety systems. Integrated safety systems coordinate several safety systems. A system coordinating several FCX-systems is therefore a special case of an integrated safety system.

In contrast to evasion assist. Evasion assist systems adapt the yaw rate of the ego-vehicle either in order to avoid a collision with an obstacle or at least to mitigate the severity of an accident.

In addition, the systems can decelerate the vehicle. Evasion assist systems can be seen as a special class of FCA-systems. In this book, they are discussed in another chapter (► Chap. 29).

In contrast to conventional assistance systems for longitudinal guidance. FCX-systems rely on machine perception systems for the outside world. That way they can be distinguished from conventional assistance systems for longitudinal control like a hydraulic brake assist system (► Sect. 1.1).

To sum up, FCX-systems use sensors for machine perception of the environment mainly designed for other systems – like the radar sensor of the ACC system. With these sensors, they avoid forward collisions or at least mitigate the impact of the accidents. Pedestrians, however, are neither explicitly represented nor recognized in state-of-the-art FCX-systems as their machine perception systems are not specialized for passenger protection and thus dedicated to the recognition of other vehicles.

4 Design Parameters of Actual Realizations

As FCX-systems are to reduce the likelihood of a forward collision, and as a result increase the safety of the passengers, they address situations in which a driver is potentially threatened with being involved in an accident. FCX-systems intervene if there is an increased likelihood that unassisted driving will end in a collision.

A reliable automatic situation assessment is crucial for the adequate intervention of FCX-systems. In this context, the item “situation” means that objects are not only described by spatial and temporal representations but also by their meaning for the ego-vehicle and its goals. For robust situation assessment, the FCX-system needs reliable recognition of the relevant objects in the driving environment by using machine perception as well as an unambiguous perception of the drivers’ intention.

The intentions of both the driver to be assisted and the drivers of other vehicles are relevant for a proper situation assessment.

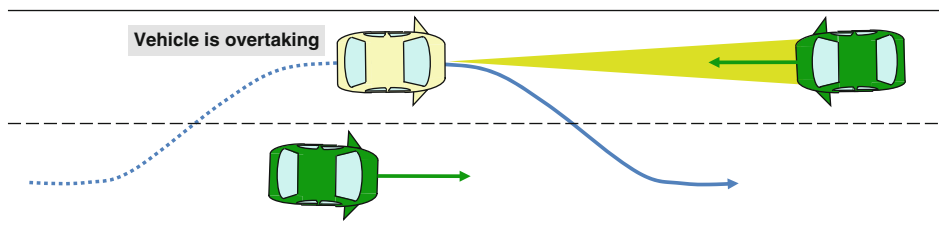
In the general case, these requirements exceed state-of-the-art technology. Any machine perception system available nowadays has relevant system limitations causing wrong assessments also in series production cars unless their field of operation is strictly limited. It is also impossible with the state-of-the-art technology to recognize the drivers’ intention in all situations automatically. Again, this uncertainty is dealt with by strictly limiting the situations of interventions.

Two degrees of freedom help to design FCX-systems ready for the market with today’s automatic situation assessment: the severity of the intervention and the –limiting – definition of situations of intervention.

4.1 CU-Criterion

The so-called CU-criterion is of special importance when limiting the FCX-situations (CU: collision unavoidable). If an accident were unavoidable, even the best driver would not be able to escape from the collision. Legally, this driver is close to the “ideal driver” who drives as well as the best 2% of all drivers. If the FCX-systems trigger only when an accident is unavoidable, rash reactions can be excluded. In rare situations – also called pathological situations – these rash decisions could lead to fatal consequences.

A short reflection may illustrate the relevance of the CU-criterion: If an automatic emergency brake is triggered while the assisted car is overtaking and before a collision is



■ Fig. 25.2

“Pathological case” of an automatic emergency brake: braking during overtaking

unavoidable, then an accident could even be caused by the trigger in case the driver would have finished the overtaking safely without being decelerated by the intervention. In this example, it is assumed contrary to state-of-the-art systems that the system would react to oncoming traffic (► Fig. 25.2).

The CU-criterion influences the definition of many interventions essentially. It is still good practice to allow automatic emergency braking only if the CU-criterion is true:

An automatic emergency brake, i.e., a braking intervention with maximum deceleration is triggered if the accident cannot be avoided by any means due to the limits of handling. So the driver is given any freedom; he is only overridden by the automatic system if he cannot avoid the collision assuming perfect driving capabilities and often even perfect weather conditions (see Kopischke 2000).

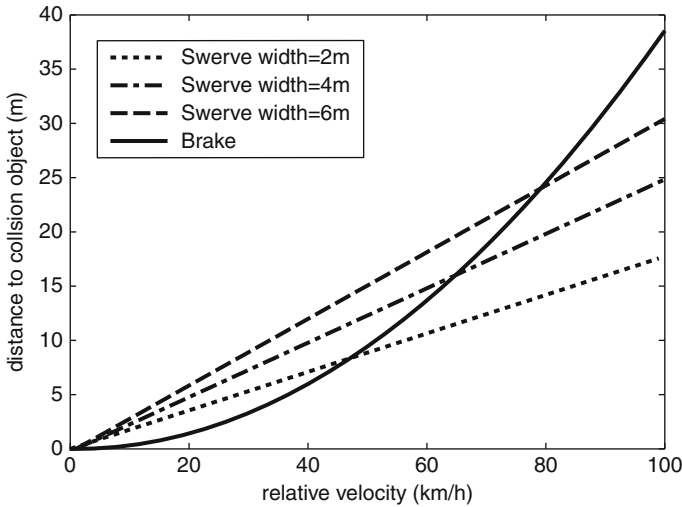
A thorough analysis shows that the CU-criterion is especially important if emergency situations with high relative velocity occur. In this case, evading by steering is still possible while braking would avoid the accident no longer. At low relative velocities the situation changes; even when evading is no longer possible, braking can still be an option to avoid the accident.

In ► Fig. 25.3, the minimal distance at which the avoiding maneuver is to be triggered is given as a function of the relative velocity. The solid black line therefore illustrates the braking distance depending on the relative velocity. The dotted line sketches the distance at which an obstacle with a width of 2 m can still be evaded by a steering maneuver. Left from the point of intersection of the two curves, the accident can be avoided by braking even if escaping by steering is no longer possible. This alters when the relative velocity increases which is shown on the right of the point of intersection.

Note that the CU-criterion discriminates whether an FCX-system is a passive safety system (FCM-system) or an active safety system (FCA-system).

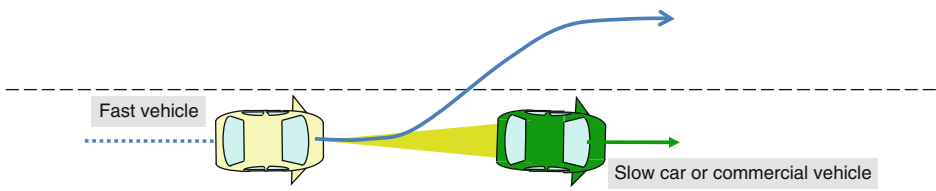
4.2 Fundamentals of Driver Warning

With respect to the limited reliability of machine perception systems currently available and to the risks of product liability linked to these limits, FCW-systems are of great importance.



■ Fig. 25.3

Influence of the relative velocity upon the CU-criterion; “width = 2 m” means that the obstacle ahead is 2 m wide



■ Fig. 25.4

Warning dilemma: approaching a slow commercial vehicle with a high relative velocity (Lucas B, 2002, Ingolstadt, personal communication)

The driver should be warned in time, so that he can still avoid the accident himself. In addition, the driver should not be irritated by false alarms. A more detailed analysis has shown that the time window for sensible warning is extremely short in the general case as the intention of the driver is not known.

A short example illustrates the challenge when interpreting the situation automatically. The driver to be assisted approaches a slowly driving commercial vehicle with a high relative velocity on the right lane of a highway with multiple lanes per direction. The lane left to the ego-vehicle (and the truck) is empty; a lane change toward this lane is both feasible and permitted. A warning in time to decelerate the vehicle behind the truck might be way too early for a sporty driver intending to change lanes shortly before contacting the commercial vehicle (► Fig. 25.4).

As this time window is so short, it is important that the driver is assisted by warnings easy to interpret and pointing intuitively to the danger ahead. Experiments show that haptic warnings by braking jerks or by a reversible belt pretensioner are very intuitive (Färber and Maurer 2005). At a braking jerk, brake pressure is shortly increased and directly afterward released again with a steep rise so that the jerk is noticeable for the passengers but the vehicle will not be slowed down significantly.

The results referred to show that the driver is made to look outside the front window but does not brake automatically. Similar reactions are reported in studies about jerking with a reversible belt pretensioner.

The example sketched above further underlines the immense importance of observing the driver's condition. Is he distracted by secondary tasks – because he is entering destinations into the navigation system or using a hands-free cellular device? Is he fatigued or is he just enjoying his dynamic style of driving being totally aware of the overall situation? Experience in the design of warning systems revealed that already relatively simple real-time warning models can significantly contribute to ease the warning dilemma (Mielich W, 2005, Ingolstadt, personal communication).

In scientific publications, two kinds of warnings are distinguished: latency warning and prior warning (e.g., Winner 2009, S. 522ff in German: “Latentwarnung” and “Vorwarnung”). A *latency warning* can be appropriate if there is no danger at all as long as the situation continues stationary. But even a minor distortion could lead to unavoidable accidents. In textbooks, the classic examples of these latent threats are vehicles following the leading vehicles within very short distances.

A *prewarning* is triggered when an accident can be predicted on the base of the current state variables.

4.3 Levels of Assistance in Dangerous Situations

Modern intervention strategies obey several partly contradicting principles:

- The vehicle should intervene in time so that the driver can avoid the accident.
- The level of assistance should be appropriate in the sense that the driver is supported but neither he nor passengers and other traffic participants are irritated by exaggerated reactions.
- The impact of interventions at the wrong time is to be minimized so that traffic safety is not endangered.

These basic principles have led to different levels of the interventions in all FCX-systems in the market. An important parameter for selecting the countermeasure is the time-to-collision (TTC) – time expected to pass until the vehicle crashes into an obstacle. Psychological experiments have revealed that also for humans the TTC is the decisive measure for situation assessment (Gibson 1950; Färber 1986).

In the following, a variety of single assistance functionalities is sketched. Each of them is part of a current realization of an FCX-system placed by various car manufacturers in their models.

4.3.1 FCC-Systems

FCC-systems condition the vehicle in such a way that in an impending dangerous situation the likelihood of the driver to survive is optimized. Actuators contributing to active and passive safety can be preconditioned accordingly.

- *Prefill*: In case of an impending FC a light pressure is established in the brake booster – experts even speak about prefilling the brakes. Thereby, the delay time is reduced as soon as the driver hits the brake pedal or another FCX-system requires brake pressure electronically (e.g., Audi A8).
- *Adaptive brake assist*: As soon as a hazardous situation is recognized by the machine perception system, the threshold defining the trigger point of the hydraulic brake assist is reduced (Zanten and Kost 2009).
- *Adaptive dampers*: At the same time, the damping parameters are adapted to reduce the stopping distance of the vehicle (e.g., Audi A8).
- *Reversible seat belt pretensioner*: The slack of the seat belt is reduced by a reversible seat belt pretensioner after fastening the belt. A further reduction of the slack is realized shortly before the crash (e.g., Audi A8).
- *Preset of the airbag*: Preset functions help to speed up the decision-making process between crash and no-crash situations based upon machine perception. The additional information is taken into account at the time of collision (Mäkinen et al. 2007; e.g., Audi A8).

4.3.2 FCW-Systems

FCW-systems warn the driver so that he is able to perceive the hazard and to prevent any accident. The severity of the warning depends on how much time is left for the driver to avoid the accident. To warn exactly in time, many systems analyze the actions of the driver either by directly observing them visually (e.g., Lexus) or by interpreting his style of driving based on current driving data.

- *Optical warning*: Warnings can be symbolic and text messages signaled by warning lamps or displays. Optical latency warnings may arise at very close distances to leading vehicles. Optical and acoustic prewarning support when the reaction time to the leading vehicles is shorter than a given threshold.
- *Acoustic warning*: Gongs, buzzers, or other sounds are meant to direct the driver's attention to dangerous situations.

- *Braking jerk*: Significant short-period changes of the actual acceleration by introducing a short pressure pulse to the brakes system are known to be a powerful warning means.
- *Reversible belt pretensioner*: Warning jerks from the reversible belt pretensioner have been proven as appropriate warning devices.
- *Active gas pedal*: As long as the driver hits the gas pedal, a growing resistance of this pedal can signal deceleration needs to the driver (e.g., Infinity).

4.3.3 Rear Impact Countermeasures

Automatic emergency braking is not appropriate to avoid all accidents or to mitigate in all traffic situations. Consider the following scenario: A forward collision with a lightweight traffic participant could be avoided by an automatic emergency braking, but a much heavier vehicle coming from behind and unable to decelerate fast enough crashes into the rear of the assisted vehicle. This is why FCX-functionalities are often accomplished by systems protecting from the following traffic.

- *Hazard warning lights*: It is almost standard that the hazard warning lights are activated if the vehicle brakes with full capacity (automatically).
- *Rearward-looking sensors*: More sophisticated FCX-systems exploit additional rearward-looking sensors to analyze whether the damage of an automatic emergency braking might exceed the benefits.

4.3.4 FCM-Systems

FCM-systems use actuators in the vehicles to reduce the severity of an impending collision.

- *Automatic partial braking*: Partial brakings are applied automatically to reduce the relative velocity and to warn the driver more drastically. They are applied in different strengths when the accident is still avoidable (e.g. Audi A8 step 1: 3 m/s^2 , step 2: 5 m/s^2).
- *CMS-braking*: CMS (Collision Mitigation Systems) were introduced in Japan starting from 2003. After a mandatory warning, the CMS-braking is activated if the driver cannot avoid an accident with a 1.0 g deceleration. The CMS-system must decelerate at least 0.5 g (e.g., Honda CMBS system, Bishop 2005).
- *Automatic emergency braking (CM)*: An automatic emergency braking is triggered if an accident is unavoidable. Depending on the current friction coefficient, automatic decelerations can reach up to 1.0 g .
- *Reversible seat belt pretensioner*: Shortly before an unavoidable crash, the seat belt is tensioned by the reversible seat belt pretensioner in order to bring the driver and the passenger into an upright position and to avoid “submarining” (Mäkinen et al. 2007).

- *Closing windows and sunroof*: Once a dangerous situation is detected, the windows and the sunroof are closed automatically. This functionality was introduced first by Daimler as part of the pre-safe system (Pre-safe: first version in 2002, Schmid et al. 2005).
- *Seat adjustments*: As part of the pre-safe system, the position of the passenger is influenced by adjusting the seats as well (Schmid et al. 2005).

4.3.5 FCA-Systems

FCA-systems try to avoid an accident.

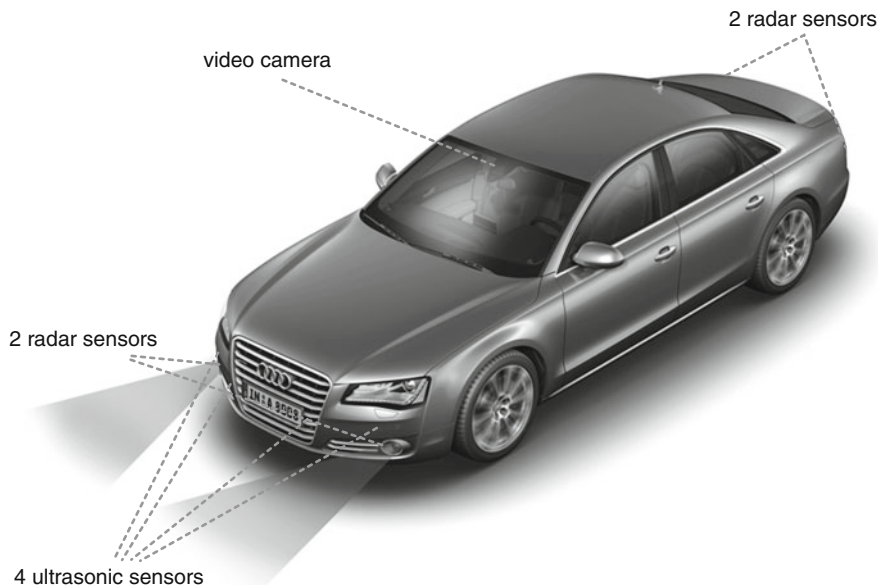
- *Target braking*: A target braking is an extension of the hydraulic brake assist. In dangerous situations, the braking driver is assisted by additional brake pressure supplied automatically in order to avoid the accident.
- *Automatic emergency braking (CA)*: An automatic emergency braking is triggered right in time to avoid an accident. Depending on the current friction coefficient, automatic decelerations can reach up to 1.0 g.
- *Other systems*: see 🔗 Chaps. 23 and 🔗 29

5 Levels of Assistance in an Actual Realization

The systems in the current Audi A8 – brand names “Braking Guard” and “PreSense” – are sketched as examples of FCX-systems. These systems reflect the current state of the art. They are adapted to other vehicles within the VW group as well (Bentley, Audi A6, Audi A7, VW Touareg). With friendly permission of the Audi AG further background information can be supplied (🔗 Sect. 7.2).

The vehicles exploit data from two forward-looking radar systems for the look ahead range (Bosch, ACC 3rd generation, 77 GHz), a monocular video camera (Bosch, 2nd generation), and ultra sonic sensors. Rearward-looking radar sensors are used to monitor the following traffic (🔗 Sect. 4.3.3) supplying mainly data to the lane change assist system (Hella, 24 GHz, 2nd generation, 🔗 Fig. 25.5).

FCX-systems are activated in several steps. In the first phase, the brakes and the dampers are preconditioned (Prefill, hydraulic brake assist, adaptive dampers, 🔗 Sect. 4.3.1). At the escalation level the driver is warned – first acoustically and optically, then by a warning jerk. In parallel, the reversible seat belt pretensioner reduces the slack of the belts of the driver and the codriver. If the driver still does not react appropriately, the automatic partial braking (first -3 m/s^2 then -5 m/s^2) and, after reaching the CU-criterion, the automatic emergency braking are activated in rapid succession. In addition, the sunroof and the windows are closed automatically; the reversible belt pretensioner increases the tightening force again. The automatic braking is signaled to the following traffic by hazard warning lights activated automatically as soon as the stronger partial braking has been triggered (🔗 Fig. 25.6).



■ Fig. 25.5

Sensors for environmental machine perception in the Audi A8 (Duba G-P, 2010, Ingolstadt, personal communication)

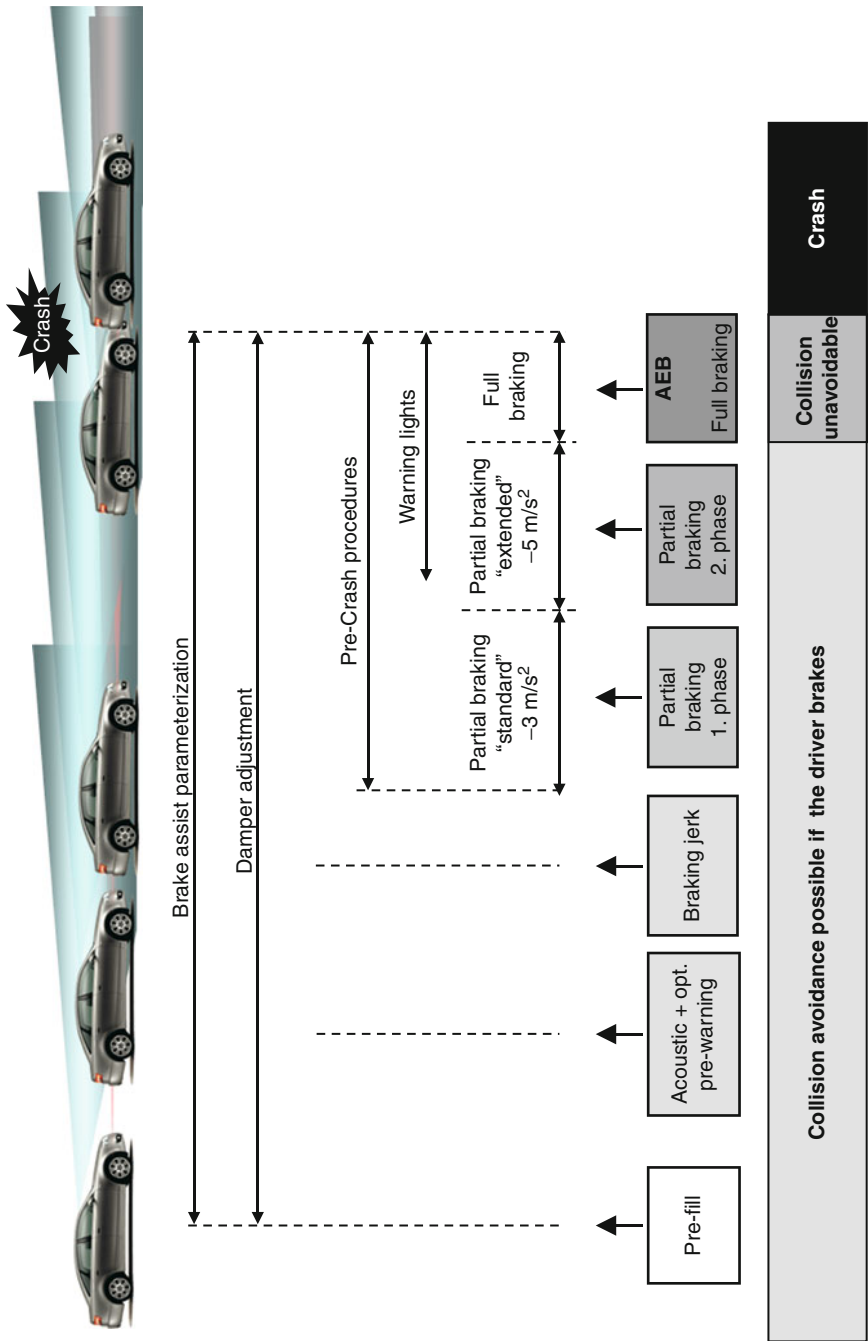
5.1 Survey of Different Realizations by the Car Manufacturers

As a special service to the respected readers, a survey of different actual realizations by the car manufacturers is offered and continuously updated at the homepage of uni-das (www.uni-das.de).

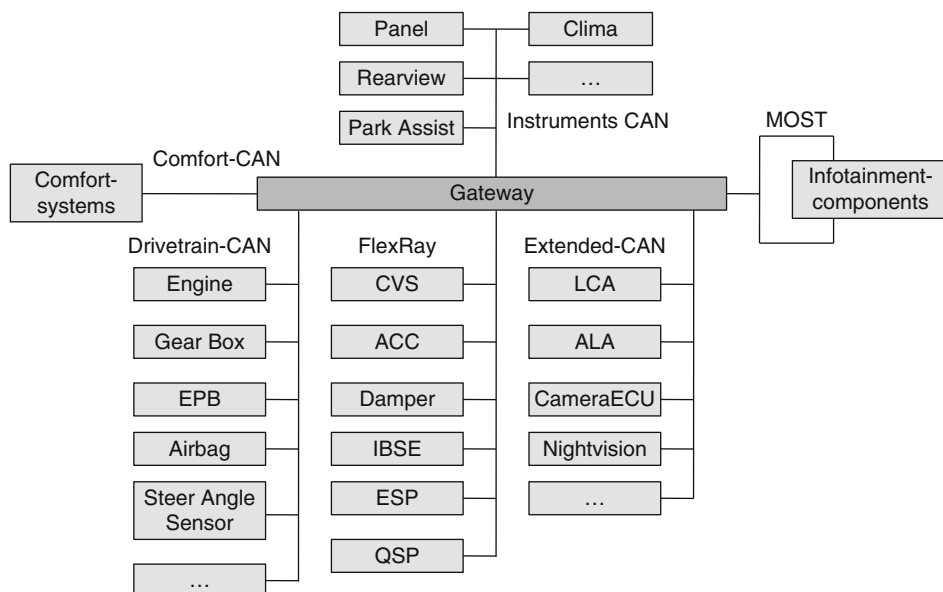
6 System Architecture

The demand on redundant multimodal environmental sensors leads to big data streams in the communication systems of modern vehicles. Availability, reliability, and the overall system safety require adequate communication technology. Time-triggered transmission and architectures in electronic control units are helpful when fusing sensor data, but eventually also for the precise control of innovative actuator systems (e.g., smart airbags).

Therefore, system architecture and its thorough planning are key factors for mastering the complexity of connected safety systems. Advisably, machine perception is taken into account already in the planning phase of the topology of the vehicle's communication systems. Data streams occurring at the fusion of sensor data derived from environmental perception can determine the topology of in-vehicle communication systems. As an example, ● Fig. 25.7 illustrates the electronic hardware architecture of the current



■ Fig. 25.6 Levels of escalation in the Audi A8 (Duba G-P, 2010, Ingolstadt, personal communication); AEB automatic emergency braking



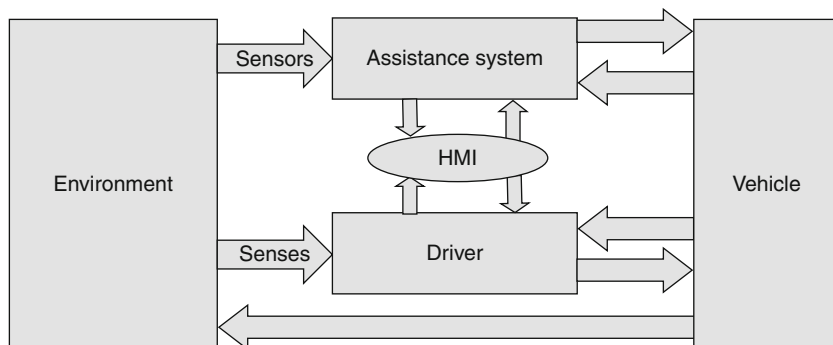
■ Fig. 25.7

Electronic hardware architecture of an Audi A8 (Kötz J, 2010, Ingolstadt, personal communication); EPB electrical parking brake, CVS computer vision system, ACC adaptive cruise control, IBSE inertial based state estimation, ESP electronic stability program, QSP Quattro sport, LCA lane change assist, ALA adaptive light assist

Audi A8 (Kötz J, 2010, Ingolstadt, personal communication). A central gateway connects several CANs, a MOST-bus for multimedia systems and a Flexray cluster for driver assistance and FCX-systems. The latter high-speed cluster couples central ECUs for computer vision (CVS: Computer Vision System), ACC, an ECU for damper control, a specialized ECU for Inertial Based State Estimation (IBSE), ESP, and QSP (Quattro Sport). All of them have to precisely interact together for proper FCX-functionalities.

Up to now the system architecture has been discussed in a very traditional way including ECUs, networks, and gateways. It has to be claimed that other aspects of the system architecture should be taken into account in order to handle the complexity in modern vehicles.

The functional system architecture splits the overall system into their functional modules. It exploits ways of representation from the fields of system dynamics and control theory (Maurer 2000). In addition, an explicit knowledge representation should be designed in order to centrally represent the overall health state of the vehicles. Up to now, this knowledge is mainly hidden in decentralized diagnosis modules. Regarding the growing autonomy of the vehicles, the knowledge of the current state is mandatory for proper vehicle reactions (Maurer 2000).



■ Fig. 25.8

Simplified block diagram of the system driver-vehicle-environment-driver assistance system (Kopf 2005, in German, modified)

Further, the properties of the vehicles should be described from the perspective of the customers and independent of their technical realization to start with (e.g., Kohoutek et al. 2007).

These three aspects can be discussed independent of the hardware; they stay untouched when migrating to another hardware platform. The hardware itself and aspects of low-level programming belong to the hardware-dependent aspects of the system architecture.

In the framework of AUTOSAR, also the low-level software is increasingly standardized. So, even software aspects can be discussed independent of the actual realization.

It can be expected that the meaning of these different aspects of the system architecture will grow significantly within the car manufacturers and their partners in order to handle the system complexity of future vehicles.

6.1 Functional System Architecture

The functional system architecture discusses the structure of the system independent of the hardware. It has been observed that there is an ongoing change of the hardware architecture in the different phases of development (research phase, concept phase, predevelopment phase) driven by the fast progress of new technologies.

In comparison, functional system architectures will only be changed if functional extensions are necessary, or completely redesigned if major paradigms are revolutionized. The functional system architecture enables hardware-independent interface analysis; it reveals how appropriate interfaces between two modules can be designed. The topology of the hardware should be derived from thorough analyses of the functional system architecture.

➤ *Figure 25.8* sketches a simple block diagram of the system driver-vehicle-environment-driver assistance system. In the system architecture, a parallel structure results because the driver and the assistance systems run the same task in parallel by definition (see Kraiss 1998).

Both the driver and the assistance system observe the environment and the ego-vehicle with their senses and the technical sensors, respectively. They influence the vehicle with appropriate actuators according to their goals. Driver and assistance system communicate via human-machine interface.

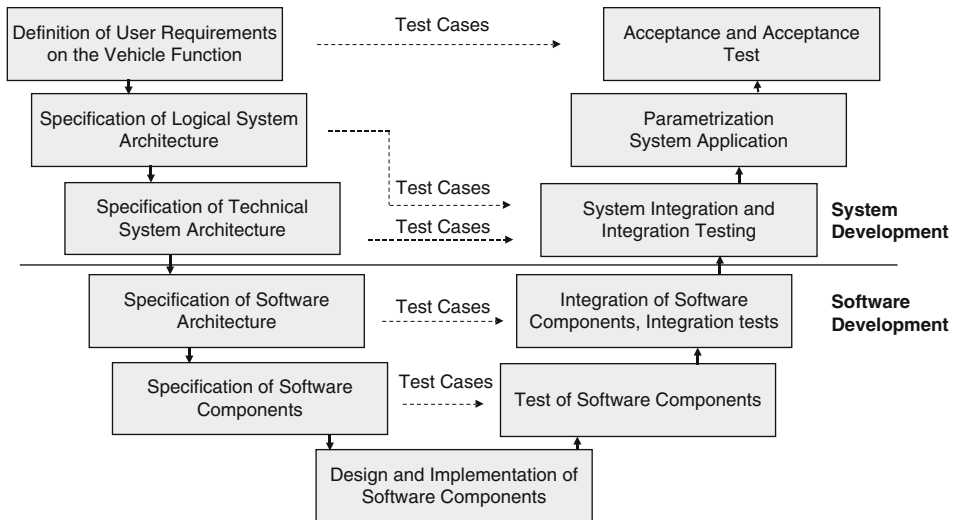
7 Design Process

7.1 Systematic Design of Driver Assistance Systems

Many developments and many design tools have been driven by military purposes. In the field of complex technical systems, the so-called V-model originally developed for defense systems has established a basic pattern for many other design schemes.

The V-model supports different fundamental design principles helping to structure complex systems. First of all, it supports the top-down design from overall requirements on the system level stepwise down to the detailed requirements on the component level. In the V-model, it is very important to specify appropriate test cases for each requirement. Corresponding to the top-down structure of the requirements, a bottom-up structure of the test cases occurs (► Fig. 25.9).

The introduction of the V-model as a paradigm in the design process of electronic vehicle systems leads to a significantly more structured way of development with car manufacturers and their system partners (e.g., Breu et al. 2007). The more the requirements are specified in detail, the more obvious it becomes that the test coverage of complex assistance systems is limited.



■ Fig. 25.9

The V-model as a state-of-the-art software design process

It is discussed critically in scientific publications that the V-model may not be appropriate if the information base is not yet complete at the beginning of the design process and therefore the system cannot be developed top-down (e.g., Reif 2006). In reality, the design proceeds incrementally and iteratively; many steps of the V-model or even the whole V-model are processed several times (Schäuffele and Zurawka 2006).

A simple design model, which was developed during the research project “Automatic Emergency Braking” at Audi, takes the need for iterative design loops into account (Maurer and Wörsdörfer 2002). The process was visualized in a simple diagram: **Figure 25.10** shows a full circle containing a complete iteration loop. An abbreviation path is defined after less than half of the circle leading back to the starting point of the development process. A more technical form of the notation was presented in 2006, but not continued during the last years (Glaser 2006). As a result, two iterative loops emerge from the structure described above: The first loop is much shorter and saves resources; it requires expert knowledge from different departments. The tasks are performed either theoretically or in more advanced labs supported by a chain of concatenated model-, software-, and/or vehicle-in-the-loop tools (Bock 2009); no prototypes are built up during the inner iterative loop. The approach is extremely powerful if the experts available within the car manufacturer – supported by external experts if necessary – identify the basic design conflicts within the inner iterative loop; in addition, they profoundly distinguish between realizable assistance functionalities on the one hand and desirable, but non-realizable ones on the other hand.

Prototypic systems will only be built up if the experts agree to a function definition as a preliminary result solving all design conflicts revealed during the theoretical discussion. Sometimes experimental setups may be required even in the iterative loop to solve basic questions.

The needs of the driver and the assistance to him are always the starting point of the design process. This may sound trivial. However, the reader interested in design of automobiles will immediately recall many examples where the driver needs were not in

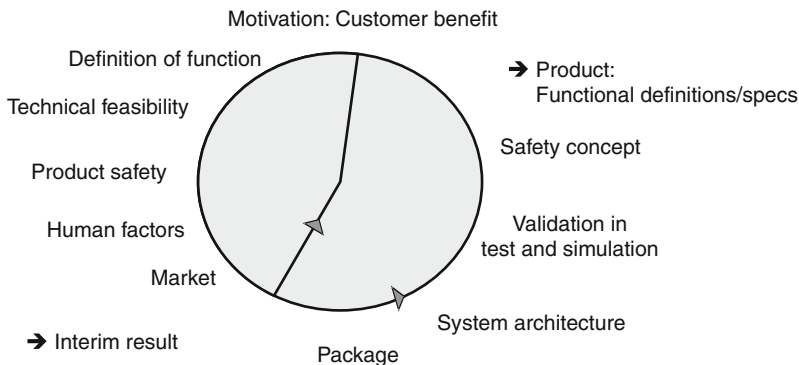


Fig. 25.10
Systematic design of driver assistance systems (Maurer and Wörsdörfer 2002, in German)

the focus of the system design. Newspapers and journals are full of examples (e.g., see Bloch 2007). Note that for the commercial success of the system the subjective driver needs, not the objective impact of the system, can make the difference. A big Japanese car manufacturer introduced one of his premium sedans equipped with drivetrain and driver assistance systems: They were effective as publicity and supported the sells, but the contributions to the driver are questionable from an expert point of view. Ideas should be derived from the identified assistance need and support the driver in technically describable scenarios.


Based on these ideas, possible assistance functions are derived and tested by the expert whether they can be realized with state-of-the-art technology. Can the functional gaps and the systems failures be controlled by untrained users in each situation? Is a user-transparent design of the assistance function and its limitations possible? Are there any sensible human-machine interfaces? Is the system to be designed financially affordable for the customer? Does it fit the branding of the car manufacturer? A more detailed discussion will include a practical example and a full iterative loop in the following section.

The approach described above supplements the class of design processes collected in the field of integrated product development (e.g., Ehrlenspiel 2003). This design scheme should be taken into account in any research and development phase of a system. User-centered and holistic design should be mandatory in academic research. In the phase of industrial research and predevelopment, the design processes are important for the commercial success of the manufacturer. The fine adjustment is performed during the series development phase especially if innovative machine perception is involved; prototypes of the sensors only available shortly before market introduction reveal whether the specification gathered at the beginning of the project will be met by real production type sensors. If they do not fulfill the specifications, it may become necessary to adjust the functionality shortly before start of production by adding yet another loop in the design scheme.

Of course, there should be open research and predevelopment projects, not directly addressed to particular user needs. But it is important that these projects are declared accordingly and do not suggest specific customer benefits.

7.2 Example: Systematic Design of an “Automatic Emergency Brake”

7.2.1 User-Oriented Definition of the Function

Analyses of accident research have revealed that many drivers do not exploit the full deceleration capacity of their vehicles. In  Fig. 25.1, the statistical evaluation of a database for accident research was displayed. Remember that in the tests a significant percentage of the driver either did not brake at all or applied a comfort braking although an emergency braking would have been the appropriate reaction to avoid the accident or at least to mitigate its impact.

Based upon these identified assistance needs, a first function is defined to start the conceptual development phase.

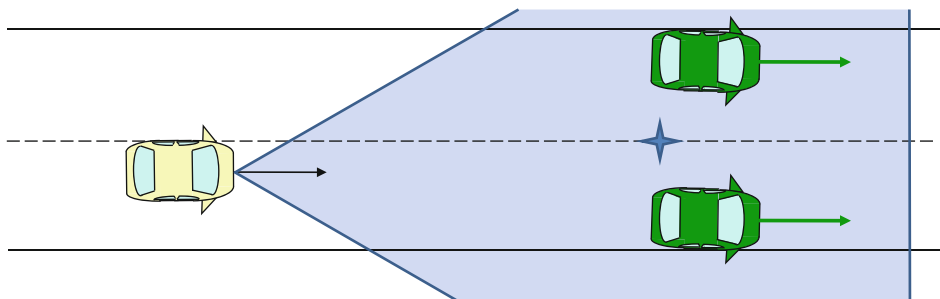
At this point, the definition explained above arises for the first time: An automatic emergency braking, i.e., a braking intervention with maximum deceleration is triggered if the accident cannot be avoided by any means due to the limits of handling. So the driver is given any freedom; he is only overridden by the automatic system if he cannot avoid the collision assuming perfect driving capabilities and often even perfect weather conditions (CU-criterion, ► Sect. 4.1).

The definition of the function incorporates the knowledge available at the beginning of this conceptual development phase: From the beginning, the system functionality is limited to accident mitigation to avoid later liability claims of drivers and their relatives who could otherwise argue that the automatic emergency brake had been triggered too early and had even caused the accident (CU-criterion).

During the first design loop the experts report that radar, lidar, or video sensors are available to realize this function as long as the scenarios are easy to describe and the weather conditions within a certain specification. As a constraint, it has to be analyzed during this first loop of iteration whether the function can be realized just by a radar sensor of a conventional ACC system. Latest during the first risk analysis, it becomes obvious that there are many possible traffic situations which exceed every single sensor principle. Missing triggers of the automatic emergency brake are regarded less critically; in this case, the assisted vehicle is not less safe than a conventional vehicle.

In contrast, an event is considered hazardous if an automatic emergency brake is activated although the CU-criterion is not fulfilled. As the fundamental principles of the single sensors are known, it is obvious to the experts that unintended activation of the automatic emergency brake may be rare, but not impossible due to the current state of the art.

Radar experts are familiar with situations in which “ghost” objects occur. That means the sensor reports objects which do even not exist. For example, this can be the case if two vehicles move with a very similar velocity and are interpreted as a single “ghost” vehicle driving in between the two real vehicles (► Fig. 25.11). It is easy to imagine how a “ghost” object could cause an erroneous reaction of the automatic emergency brake.



■ Fig. 25.11
“Ghost” objects perceived by a radar system

In the design discussions, it will be taken seriously when experts in product liability argue that in the event of a failure courts will look for analogies. A likely analogy will be taken from machine perception where important state variables are measured or derived redundantly in a vehicle dynamics control system. Therefore, the automatic emergency brake has to be designed with redundant means for perceiving the decisive parameters.

Even in this very early phase, experts underline that the expected limits of the function should be comprehensible for the driver and the other traffic participants. They emphasize that the manufacturer is responsible for the expectations of the customer. This expert knowledge is incorporated in variant design aids thanks to the projects in the framework of response (e.g., Kopf 1999 and Donner et al. 2007).

As an additional requirement the system has to monitor itself, realize any substantial degradation, and warn the driver adequately.

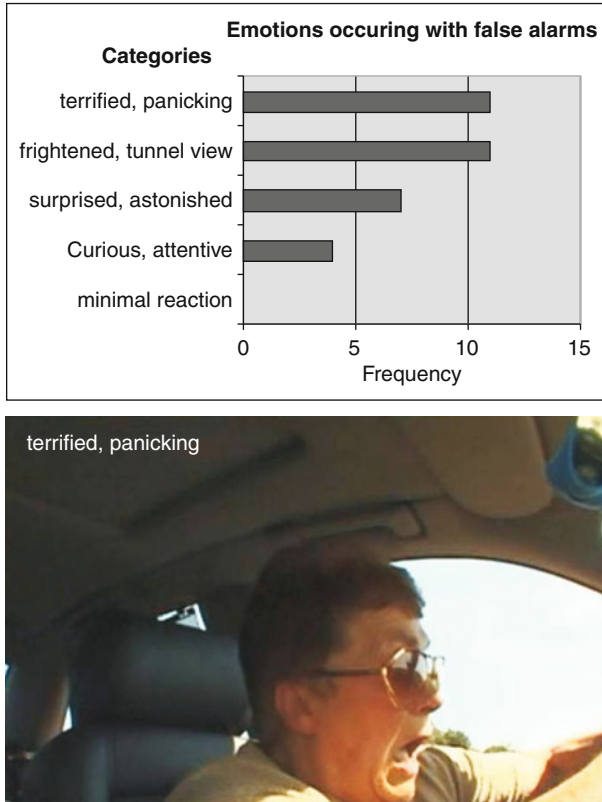
In order to prove that the system has worked without any malfunctions, data recorders are regarded as a sensible extension of advanced driver assistance systems.

In the discussion of possible hazards, it is crucial whether false triggers of an automatic emergency brake are controllable for the driver and the following traffic. The thorough analysis of this question requires building prototypes for the first time in the concept phase and therefore a first iteration of the outer design loop. The test results are unambiguous: More than a third of the drivers' reactions are categorized as "terrified, panicking." Another third of the drivers react "frightened, with tunnel view." It cannot be excluded that the comparably mild ("surprised" or "curious") reactions were evoked because the experiment was performed at a test track closed for public traffic (► Fig. 25.12, Färber and Maurer 2005).

These analyses underline that false alarms of an automatic emergency brake can cause a significant risk for the driver, the following traffic, the car manufacturer, and the system partner. In addition to the technical, ergonomic, and legal aspects, product marketing should be addressed already during the concept phase. How do costly technical innovations contribute to the manufacturer if they do not fit into the image of the brand? They will not be promoted by the manufacturer and bought by the customers. If it comes to assistance functions, the situation is even more complicated: As a result of the expected functional limits the products will not be promoted aggressively anyway. The manufacturer is responsible for the customers' expectations.

From the first iterative design loop the following is learnt: A functional definition has been identified with a big potential impact to traffic safety. The sensors specified in the development task limit the benefits to longitudinal traffic. The realization with ACC-sensors already launched in the market would be inexpensive. Comparisons with other safety systems reveal that a redundant perception of the most relevant state variables is strictly required. Experimental analyses which had to be performed during this early stage of development underlined that false triggers of the functional definition mentioned above are not acceptable.

As a consistent preliminary result was not found during the first loop, the further development has to be modified fundamentally. Within a longer-term perspective, the false alarm rate should be minimized by perception systems as complementary as possible.



■ Fig. 25.12

Drivers' reactions to an unjustified trigger of an automatic emergency brake (Färber and Maurer 2005, in German)

In the short term, a consistent function should be reached by varying the functional definition.

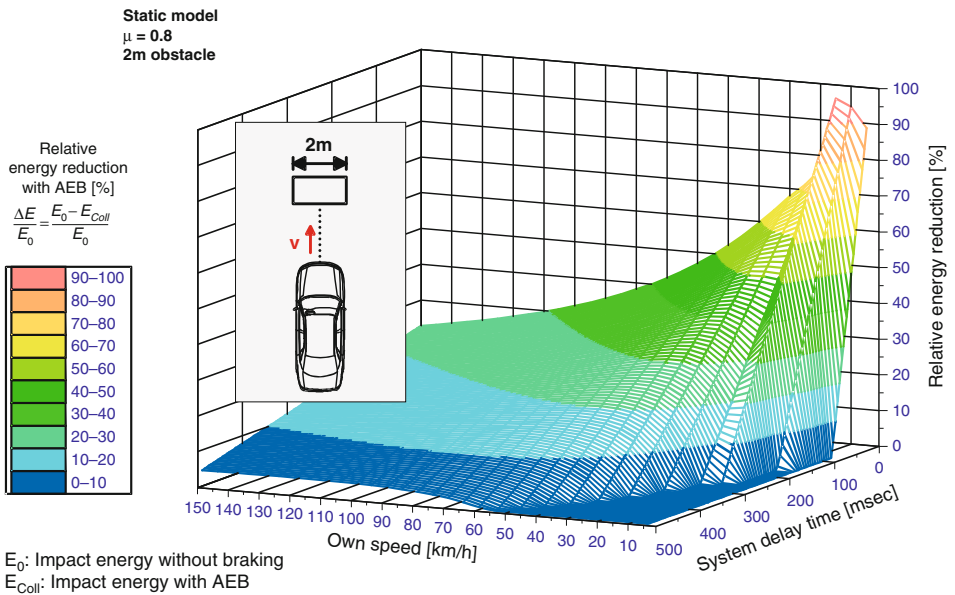
In the tests, the effect of the false emergency braking was impressive. Could not a weak braking jerk, e.g., a simple warning with a haptic jerk, be an alternative way to point the driver to a hazard ahead? In case of a mistimed jerk, the following traffic would then not be endangered by a sudden unexpected deceleration of the vehicle. Experimental analyses confirm both expectations. The warning jerk is an efficient warning device; with an appropriate braking system, there is no noticeable deceleration.

In the next iterative loop, a warning system is therefore matured for the market assisting the driver as described above. As the intervention is uncritical, even if it is unjustified, a false alarm rate of 1 per 10,000 km seems to be acceptable. This time the preliminary results look promising: The warning via the haptic channel is both direct and effective. Thus, a significant customer benefit is predicted. The impact again is limited to longitudinal traffic as the system design is based upon ACC-sensors. The function can be realized without any additional sensor hardware. False alarms are both controllable

and accepted by the drivers. This function can be offered to the market shortly after the concept phase (product name: Audi: “Audi Braking Guard”; VW: “Front Scan,” market introduction: 2006).

The further development of the original idea of an automatic emergency brake requires more sophisticated solutions. Quantitative prognoses can be given for the functional definition developed during the first design loop. The analysis which parameters influence the benefits most has also great importance. ● Figure 25.13 shows how the relative energy reduction and, therefore, the impact on accident mitigation depend on the system delay. This representation may even be helpful to communicate the benefit of a faster brake system within the car manufacturer or to select sensors which fulfill the requirements on the dynamics (Kopischke S, 2000, Ingolstadt, personal communication quoted by Maurer 2009).

During the first design loop, it had become apparent that crucial state variables are to be perceived redundantly. The selection of an appropriate sensor configuration is one of the most challenging design tasks when developing an innovative driver assistance system. In the general case, appropriate sensors which fulfill all requirements derived from the functionality are not available on the market. Even measures to compare different sensor set and perception algorithms have not yet been established. So the selection of the sensor set will be based on the performance of the current prototypes and the performance predicted by the developers for their sensors in series production.



■ Fig. 25.13

Sensitivity of the relative energy reduction of an automatic emergency braking to the delay time (Kopischke S, 2000, Ingolstadt, personal communication)

Apart from this uncertainty, the robustness of the machine perception can be increased by combining appropriate sensor principles. To establish both robustness of the perception system and formal redundancy, combinations of many different sensor principles are taken into account. A far-range radar sensor is the preferred sensor of the developers responsible for the design of adaptive cruise control (see 🔗 Chap. 24) due to its performance in adverse weather conditions. A mono-video camera is becoming standard for lane departure warning and traffic sign recognition systems. At the point of decision, it was not sure whether additional redundancy would be needed. Therefore, stereo vision, laser sensors, and photonic mixing device (PMD) were analyzed as an additional forward-looking sensor. Finally, the sensor system consists of two radar sensors and a video camera (Lucas et al. 2008; Duba G-P, 2010, Ingolstadt, personal communication).

The reliability of this highly sophisticated sensor set is significantly higher compared to a single ACC radar sensor. In order to further reduce the likelihood of dangerous false interventions of an automatic emergency brake, the data of the rearward-looking radar system are exploited as well. An emergency braking is only triggered with full braking power if there is no vehicle following the ego-vehicle closely.

The benefit of the system is again limited to longitudinal traffic. The likelihood of false alarms is minimized. In this third iterative loop, a few aspects from the outer loop are focused on, too: packaging and functional tests. Aspects of the system architecture have been treated in another section (🔗 Sect. 6).

Outside the car manufacturer, the integration of the sensors in the design concept is widely underestimated. Package spaces are threatened in any car especially if needed for other functions than the basic functions of the vehicle; this is even true if they have been reserved in the early concept phase. The integration of the ACC-sensors in the Audi A8 is a real highlight: They overtook the previous package space of the fog lights now integrated into the headlights. Is this not visible progress in car design: radar sensors replacing (old-fashioned) fog lights?

7.2.2 Functional Tests of Driver Assistance Systems

Nowadays the usage of the term “testing” is not specific in the automobile development. It encompasses different test categories like functional tests, usability tests, user transparency tests, customer acceptance tests, electromagnetic interference tests, climate stress tests, acoustic tests, crash tests, electric and electronic tests including hardware-in-the-loop tests, software integrity tests. The list can easily be extended. In the field of driver assistance, each topic itself is challenging, complex, and worth its own chapter in this book. Many groups in the technical development and in the quality assurance department are in charge of and contribute to the testing process.

In this chapter, the focus is on the aspects of functional tests again illustrated by the design example of an automatic emergency brake. Two failures are of outstanding importance because they decide on how the system is perceived by the driver and the public. It has already been reported that the drivers’ reactions to unjustified interventions

were significant. So they have to be avoided by all means. The experimental setup for tests of justified interventions is even more challenging as they will always lead to a crash according to the functional definition.

It is still an open question which error rate would be accepted in society. The standard IEC 61508 specified acceptable error rates depending on the safety integrity level (SIL). As this standard is both detailed and modified by the ISO standard 26262, fundamental design principles tend to replace strict error rates. An open discussion with all groups including car manufacturers, insurance companies, motor clubs, the government represented by departments and agencies, and related nongovernmental organizations would be helpful to back up the activities of the automobile industry. Homann (2005) sketched the outline of a discussion in an enlightened state referring to examples from aviation and medicine.

7.2.3 Test Case “Justified Intervention”: Vehicle-in-the-Loop

A few requirements for testing the test case “justified intervention” are given here:

- An automatic emergency brake will be triggered.
- There will be a crash.
- The driver and the vehicle must not be exposed to danger.
- The situations should look realistically to the driver.
- The test should be as reproducible as possible.

Simple test setups or examinations in the driving simulator do not fulfill all criteria. In the driving simulator, the threat to the driver may be not as realistic as necessary; the dynamics of the vehicle are limited in the simulator. If the real vehicle crashes into foam cubes, jibs, and small mobile vehicles, again the driver will not panic at all. The most advanced setup is reported by (Hurich et al. 2009) in which FCX-systems are challenged by real autonomous cars. Of course, this very expensive setup does not yet include intentional crashes.

All the requirements mentioned above are met by a new testing method for driver assistance systems called vehicle-in-the-loop (VIL, ► Fig. 25.14, Bock et al. 2007). The basic idea is that only the other traffic participants are simulated. Everything else is real: A real driver conducts a real vehicle on a real test track. The reality is augmented via see-through glasses displaying the simulated traffic participants to the driver. Experiments show that the driver reacts very realistically to this simulation even though he can distinguish between real and simulated objects when asked.

7.2.4 Error Probability for “Unjustified Interventions”: Trojan Horses

It is as challenging as the test of the justified intervention described above to ensure that the error probability rate is very little – as an example no more than $10^{-9}/h$ (according to the former IEC 61508 standard). Assuming that a vehicle only drives 30 km/h in average, this means that 30 billion test kilometers are to be driven without any false



■ Fig. 25.14
Vehicle-in-the-loop: basic principle (Bock 2009)

alarm. This vast mileage cannot be driven in standard fleet tests due to financial reasons. Again alternative testing methods are necessary. Winner (2002) proposed to implement Trojan horses in vehicles of the customers. The idea is that the customers would purchase a comfort function like ACC realized with the same sensor configuration and the same perception software. In addition, the software would contain all functions of an automatic emergency brake, but the FCM-module would not have access to the braking system. Instead of intervening the FCM-system would write an entry in the permanent memory of the ECU (electronic control unit). If entries are detected during a later service, they either result from an accident which should be known or they would have been caused by a false alarm. In principle, all information would be available to determine the wanted error probability rate. Currently, the car manufacturers do not publicly discuss whether the process is appropriate for future testing. However, it cannot be ruled out that car manufacturers or their system partners use this method already without communicating.

8 Conclusion

Forward collisions represent a significant portion of all severe accidents. This is why appropriate warning and collision avoidance systems are of great importance to increase traffic safety. Different system specifications are subsumed under the so-called FCX-systems; they differ in their way of affecting the overall system driver-vehicle-environment as forward collision conditioning, forward collision mitigation, forward collision warning, and forward collision avoidance systems.

The specifications of the actual systems already on the market can only be understood if the characteristics of machine perception are considered carefully. The progress in the field of machine perception enables the forward collision warning and avoidance systems. There are still limitations of state-of-the-art perception systems compared to attentive human drivers which must be considered when designing FCX-functions.

A systematic design process is recommended for FCX-systems: The motivation for the assistance is always to be derived from accident research. Already in the early conceptual phase functional safety, legal, ergonomic, and marketing aspects should be taken into consideration. Only if a consistent functional specification is found, further developments including package and architecture aspects are justified. Concepts for testing and evaluation should be designed in an early development phase as well.

Major challenges are to be expected when introducing even more advanced and more intervening systems. The automobile industry is prepared to design them but methods for testing autonomous interventions with even higher impacts on system dynamics are not yet established.

References

- Bainbridge L (1983) Ironies of automation. *Automatica* 19(6):775–779
- Bishop R (2005) Intelligent vehicles technology and trends. Artech House, Boston
- Bloch A (2007) Tech no. www.auto-motor-und-sport.de, 16/2007
- Bock T (2009) Bewertung von Fahrerassistenzsystemen mittels der Vehicle in the Loop-Simulation. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme, 1st edn. Vieweg und Teubner, Wiesbaden, pp 76ff
- Bock T, Maurer M, Färber B (2007) Vehicle in the Loop (VIL) – a new simulator set-up for testing advanced driving assistance systems. In: Driving simulation conference North America 2007, University of Iowa
- Breu A, Holzmann M, Maurer M, Hilgers A (2007) Prozess zur Komplexitätsbeherrschung bei der Entwicklung eines Stillstandsmanagements für ein hochvernetztes Fahrerassistenzsystem. Stillstandsmanagement. 8.-9. November 2007, Haus der Technik, Essen
- Bubb H (1981) The influence of braking distance indication on the driver's behaviour. In: Osborne DJ, Levis JA (eds) Human factors in transport research, Academic Press, London vol 1 and 2, p 338
- Buld S, Tietze H, Krüger H-P (2005) Auswirkungen von Teilautomation auf das Fahren. In: Maurer M, Stiller C (eds) Fahrerassistenzsysteme mit maschineller Wahrnehmung. Springer, Heidelberg
- Donner E, Winkle T, Walz R, Schwarz J (2007) RESPONSE 3 – Code of Practice für die Entwicklung, Validierung und Markteinführung von Fahrerassistenzsystemen. VDA. Technischer Kongress. 28–29 März 2007, Sindelfingen
- Ehrlenspiel K (2003) Integrierte Produktentwicklung. Hanser, München
- eSafety (2010) eSafetySupport. www.esafetysupport.org
- Färber B (1986) Abstandswahrnehmung und Bremsverhalten von Kraftfahrern im fließenden Verkehr. *Z Verkehrssicherh* 32:9–13
- Färber B, Maurer M (2005) Nutzer- und Nutzenparameter von Collision Warning und Collision Mitigation Systemen, In: Maurer M, Stiller C (eds) Workshop Fahrerassistenzsysteme FAS2005, Walting
- Gelau C, Gasser TM, Seeck A (2009) Fahrerassistenz und Verkehrssicherheit. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme, 1st edn. Vieweg und Teubner, Wiesbaden, p 26f
- Gibson JJ (1950) The perception of the visual world. Houghton Mifflin, Cambridge, MA
- Glaser H (2006) Fahrwerk und Fahrerassistenz – eine ideale Kombination? In: 7. Symposium Automatisierungs-, Assistenzsysteme und eingebettete Systeme für Transportmittel. AAET 2006, 21–23 Feb 2006, Braunschweig
- Homann K (2005) Wirtschaft und gesellschaftliche Akzeptanz: Fahrerassistenzsysteme auf dem Prüfstand. In: Maurer M, Stiller C (eds) Fahrerassistenzsysteme mit maschineller Wahrnehmung. Springer, Heidelberg
- Hurich W, Luther J, Schöner HP (2009) Koordiniertes Automatisiertes Fahren zum Entwickeln, Prüfen und Absichern von Assistenzsystemen. In: 10. Braunschweiger Symposium AAET 2009, Braunschweig, February

- Kiesewetter W, Klinkner W, Reichelt W, Steiner M (1997) Der neue Brake Assist von Mercedes Benz – aktive Fahrerunterstützung in Notsituationen. In: ATZ Automobiltechnische Zeitschrift 99 No. 6
- Kohoutek P, Dietz J, Burggraf B (2007) Entwicklungsziele und Konzeptauslegung des neuen Audi A4. In: ATZ/MTZ extra – Der neue Audi A4, September 2007, Vieweg, Wiesbaden
- Kopf M (1999) RESPONSE checklist for theoretical assessment of advanced driver assistance systems: methods, results and assessment of applicability, European Commission DG XIII: Project TR4022 Deliverable: D4.2, September 1999
- Kopf M (2005) Was nützt es dem Fahrer, wenn Fahrerinformationssysteme und –assistenzsysteme etwas über ihn wissen. In: Maurer M, Stiller C (eds) Fahrerassistenzsysteme mit maschineller Wahrnehmung. Springer, Heidelberg
- Kopischke S (2000) Entwicklung einer Notbremsfunktion mit Rapid Prototyping Methoden. Dissertation, TU Braunschweig
- Kraiss K-F (1998) Benutzergerechte Automatisierung - Grundlagen und Realisierungskonzepte. In: at – Automatisierungstechnik 46, Band 10, S. 457–467, Oldenbourg, München
- Lucas B, Held R, Duba G-P, Maurer M, Klar M, Freundt D (2008) Frontsensordsystem mit Doppel Long Range Radar. In: Maurer M, Stiller C (eds) 5. Workshop Fahrerassistenzsysteme, Walting
- Mäkinen T, Irion J, Miglietta M, Tango F, Broggi A, Bertozzi M, Appenrodt N, Hackbarth T, Nilsson J, Sjogren A, Sohnke T, Kibbel J (2007) APALACI final report 50.10b, February 2007
- Maurer M (2000) Flexible Automatisierung von Straßenfahrzeugen mit Rechnersehen. Fortschritt-Berichte VDI, Reihe 12: Verkehrstechnik/Fahrzeugtechnik. Bd. 443
- Maurer M (2009) Entwurf und Test von Fahrerassistenzsystemen. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme, 1st edn. Vieweg und Teubner, Wiesbaden
- Maurer M, Wörsdörfer K-F (2002) Unfallschwereminderung durch Fahrerassistenzsysteme mit maschineller Wahrnehmung – Potentiale und Risiken, Unterlagen zum Seminar Fahrerassistenzsysteme und aktive Sicherheit, Haus der Technik, Essen, 20 Nov 2002
- Naab K, Reichart G (1998) Grundlagen der Fahrerassistenz und Anforderungen aus Nutzersicht, Seminar Fahrerassistenzsysteme, Haus der Technik, Essen, 16–17 Nov 1998
- Neukum A, Lübbecke T, Krüger H-P, Mayser C, Steinle J (2008) ACC Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In: Maurer M, Stiller C (eds) 5. Workshop Fahrerassistenzsysteme, Walting
- Reif K (2006) Automobilelektronik – Eine Einführung für Ingenieure. ATZ/MTZ-Fachbuch, Vieweg
- Schäuffele J, Zurawka T (2006) Automotive software engineering, 3rd edn. ATZ/MTZ-Fachbuch, Vieweg
- Schmid V, Bernzen W, Schmitt J, Reutter D (2005) Eine neue Dimension der Aktiven und Passiven Sicherheit mit PRE-SAFE und Bremsassistent BAS PLUS in der neuen Mercedes-Benz S-Klasse. In: 12. Internationaler Kongress Elektronik im Kraftfahrzeug, Baden-Baden, VDI-Berichte 1907
- Weinberger M (2001) Der Einfluss von Adaptive Cruise Control Systemen auf das Fahrverhalten, Dissertation, TU München, Berichte aus der Ergonomie, Shaker-Verlag, Aachen
- Wiesbeck W (2006) Radar system engineering. Uni Karlsruhe, Lecture script, 13th edn. http://www2.ihe.uni-karlsruhe.de/lehre/grt/RSE_LectureScript_WS0607.pdf
- Winner H (2002) Einrichtung zum Bereitstellen von Signalen in einem Kraftfahrzeug. Patent DE 101 02 771 A1, Deutsches Patent- und Markenamt, Anmeldetag: 23 Jan 2001, Offenlegungstag: 25 July 2002
- Winner H (2009) Frontalkollisionsschutzsysteme. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme, 1st edn. Vieweg und Teubner, Wiesbaden
- Zanten vA, Kost F (2009) Bremsenbasierte Assistenzfunktionen. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme, 1st edn. Vieweg und Teubner, Wiesbaden, p 392f

26 Lane Departure and Lane Keeping

Jens E. Gayko

VDE Association for Electrical, Electronic and Information Technologies, VDE Headquarters, Frankfurt am Main, Germany

1	<i>Introduction</i>	690
2	<i>Function Description</i>	691
2.1	Lane Departure Warning Function	691
2.2	Lane Keeping Assist Systems	693
2.3	Combined and Integrated Lateral Support Systems	695
3	<i>Development of LDW and LKA Systems</i>	696
3.1	Lane Marker Recognition	697
3.2	Warning Elements	700
3.3	Lane Keeping Assistance Controller and Actuator (Just LKAS)	702
4	<i>Market Trend and Outlook</i>	703
5	<i>Conclusion</i>	707

Abstract: Supporting the driver for keeping the lane is the aim of lane departure warning and lane keeping systems. Unintended lane departures account for a high percentage of road accidents in general and especially of severe accidents. Therefore, these systems have been researched intensively and introduced in the market for several vehicles from heavy trucks to compact sedans. The systems described in this chapter aim at increasing the safety of driving or increasing the comfort on long journeys or a combination of both. The characteristics of the different variants are described as well as key technologies of state-of-the-art implementations.

1 Introduction

Keeping the lane is one of the primary tasks of controlling a vehicle. Especially on long trips on extra-urban roads this might be a monotonous and annoying task where unintended lane departures may occur caused by momentary lapses of attention or drowsiness. As a consequence of an unintended lane departure there might happen a

- Collision with a stationary object
- Collision with a vehicle traveling in the same direction
- Collision with oncoming traffic
- Rollover accident
- Collision with pedestrians and bicyclists beside the road
- Further accident caused by failed corrective steering and braking (loss of vehicle control)

Generally speaking unintended lane departures are one major cause of severe accidents. However, there might be various root causes of an unintended lane departure like driver distraction, drowsiness, a temporary blackout of the driver, too high velocity (especially in curves), poor visibility of lane markers and road geometry, and more. These chains of cause and effect need to be considered for defining the scope of lane keeping and lane departure warning systems as well as their effectiveness.

The systems described in this chapter aim at improving driving safety by preventing unintended lane departures. In addition some systems aim at improving comfort by releasing the driver from monotonous tasks on highways and highway-like roads. The support of drivers at roadway sections having temporary or irregular lane markings (such as roadwork zones) is not within the scope of today's available systems. Therefore, this aspect is discussed at the outlook at the end of the chapter. Further lateral support systems like curve over-speed countermeasures, lane change assist, and blind spot warning systems are not covered by this chapter. The next section describes the functional aspects of the regarded systems in more detail.

2 Function Description

From a functional point of view there are three main categories of systems to support the driver in keeping the lane and to avoid unintended drifting out of the lane:

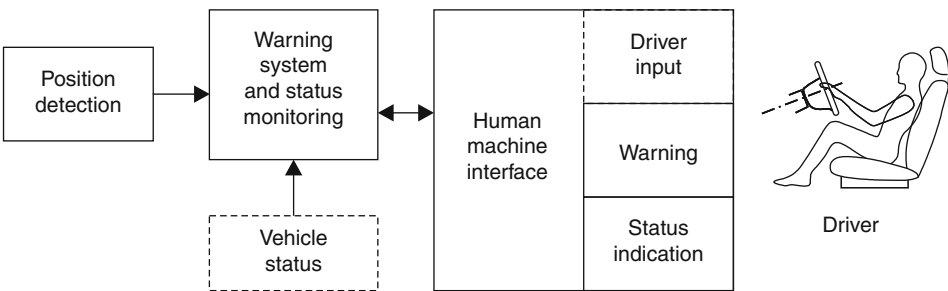
1. Automated lane keeping
2. Lane departure warning systems (LDWS)
3. Lane keeping support/assist systems (LKAS)

Automated lane keeping has been subject of academic and industrial research since decades and first test vehicles showed more than 90% of automatic driving more than 10 years ago (Pomerleau 1995; Broggi et al. 2001; Dickmanns 2007). Currently automated lane keeping is not in line with the legal framework in the major automobile regions worldwide as defined by the Vienna Convention (United Nations 1968). However, automated driving like platooning is (again) part of the research agenda in some regions. This time research is motivated by its potential to realize significant fuel savings. So maybe the legal framework will be modified in the future and thus automated lane keeping will become of relevance for series deployment in the future. Since this chapter mainly focuses on state-of-the-art systems available on the market automated lane keeping is not discussed in more detail in the following sections.

2.1 Lane Departure Warning Function

Lane departure warning systems are in-vehicle systems that warn the driver of an unintended lane departure on highways and highway-like roads, with one or two lane markings depending on the type of system. LDWS have no actors for influencing the vehicle heading. These systems provide a warning to the driver and request him to initiate proper actions to avoid any unintended maneuver. In the simplest case two warning zones are defined.

The functional elements of a LDWS are shown in [Fig. 26.1](#). The position detection unit recognizes the lane markers and enables the warning system to decide if a warning



■ Fig. 26.1

Functional elements of a lane departure warning system

should be issued. For this purpose the lateral deviation from the lane border needs to be extracted. In addition the rate of departure as well as the curvature of the upcoming road segment might be used but are not required necessarily.

Also it should be noted that LDW function can be operational if just the lane markers of one side are visible. In that case the system may assume a default lane width to establish a virtual lane border on the opposite side from the visible lane marking.

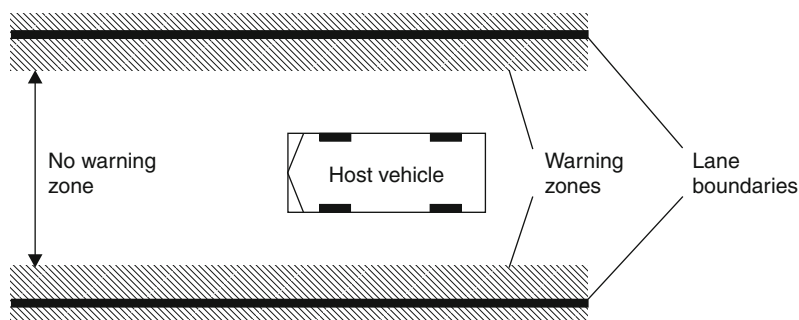
Technically position detection unit could be realized either by a camera with image processing for lane marker recognition or by a laser sensor. The warning system decides, based on the information of the position detection unit, if a warning should be issued. In the simplest case the lateral distance DLC (Distance to Line Crossing) to a lane boundary is used as criterion. In Fig. 26.2, a birds' eye view of a vehicle traveling in a lane is shown. The inner zone of the lane is the "no warning zone" whereas the warning zones cover the area of the lane boundaries. In order to cope with detection inaccuracies a "may warn" zone could be defined in between the no warning zone and the warning zones.

More elaborated warning systems use predictive criteria like "Time to Line Crossing" TLC (sometimes referred as $TTLC$) to issue a warning shortly before the lane departure happens (Van Winsum et al. 2000; Mammar et al. 2006). TLC could be defined as

$$TLC = \frac{D}{V_Y} \quad (26.1)$$

where D is the lateral distance of a defined part of the vehicle to the lane boundary and V_Y is the rate of departure for a vehicle traveling straight. This simplified formula does not consider curvatures. The geometrical conditions for curved road segments and vehicles driving on a curved trajectory can be found in the literature referenced above.

Status monitoring and status indication detect if the LDWS is operational and indicate the status to the driver. Functional aspects of status monitoring include monitoring of on/off switch (if installed) as driver input and velocity preconditions (if given) via vehicle status monitoring. The nonfunctional aspects of status monitoring are discussed in Sect. 3. Activation of LDW function is typically realized using an on/off switch which will keep the status after ignition off.



■ Fig. 26.2

Warning threshold zones and movement of vehicle inside the lane

Vehicle status monitoring can be realized using data from the in-vehicle network (e.g., CAN-bus, speed pulse signal, dedicated connection) to monitor the vehicle velocity, or detecting intended lane departures by analyzing the turn signal indicators and optionally driver's steering behavior. Alternatively vehicle velocity sensing can be realized using GPS. A basic LDW system can be implemented without vehicle status monitoring but such a system might suffer from frequent false alarms during intended lane changes (because there will be no detection of turn signal).

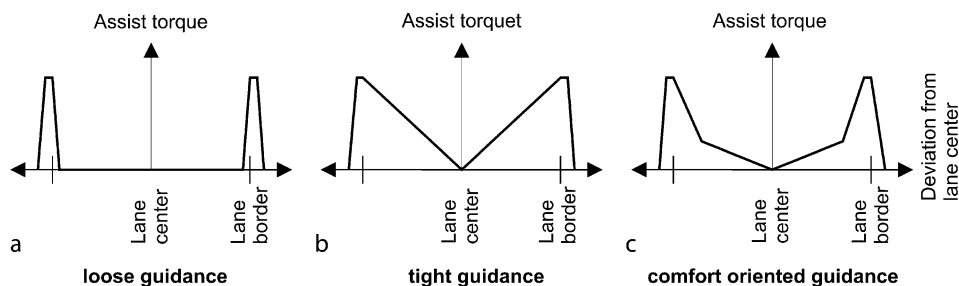
The warning element is the most important element of the human-machine interface. A proper design of the used warning element is important for the effectiveness of the LDWS but will also influence the user acceptance of the system. Warning elements might address various human senses: seeing, hearing, and feeling. Brief discussions of different warning elements will be given later in [Sect. 3.2](#).

Lane departure warning systems debuted commercially for heavy duty trucks in Europe in 2000, followed shortly thereafter by introductions in the United States. In the following years several LDWS have been introduced or announced for several vehicles from compact sedans to luxury sedans and trucks by (in alphabetical order) Audi, BMW, Citroën/Peugeot (2004 for Europe), Daimler, Fiat, Ford, GM/Opel, Hyundai/Kia (2011), (Taiwan-based manufacturer) Luxgen, Nissan (2005 for United States), Saab, Volvo (2007) in different markets worldwide. Furthermore several aftermarket systems are available. For example, the LDWS of company Mobileye is available for several brands as manufacturer integrated system as well as aftermarket solution. Aftermarket systems gained some popularity for heavy duty truck market. Recently a software solution for smartphones with integrated camera has been introduced. As a recent trend so-called multipurpose cameras are deployed featuring several additional functions like speed limit sign recognition (mainly for Europe) and detection of oncoming vehicles (for high-beam control).

2.2 Lane Keeping Assist Systems

Lane keeping assist systems support the driver in keeping the lane by actively influencing the heading of the vehicle but do not release him from the task of lane keeping. The control variable for LKA systems is the position of the vehicle within the lane. As actuating variable the majority of commercially available LKAS use an assist torque to the vehicle steering. Typically LKA systems have to be activated manually by the driver, after entering the highway. For some systems the activation of LKAS is coupled with the adaptive cruise control system. All commercially available LKA systems can be overridden by the driver at any time. This is an important property to avoid product liability issues and patronizing of the driver.

There are different variants of LKA systems which can be described by the characteristic of the actuating variable dependent on the control variable. In [Fig. 26.1](#), the assist torque is shown against the deviation from the lane center in schematic diagrams. The left diagram called “loose guidance” characterizes a system which supports the driver in case



■ Fig. 26.3

Examples of assist torque of lane keeping assist systems; (a) loose guidance, (b) tight guidance, and (c) comfort-oriented guidance

of imminent lane departure. Support is given by a corrective steering torque. Depending on the force of the assist torque driver's corrective action is required or the system is able to keep the lane without driver support. Typically these systems focus on the safety aspect of LKAS rather than the comfort aspect. Such a torque characteristic has been introduced, for example, realized by Volkswagen in 2008 in Europe and 2009 in Japan.

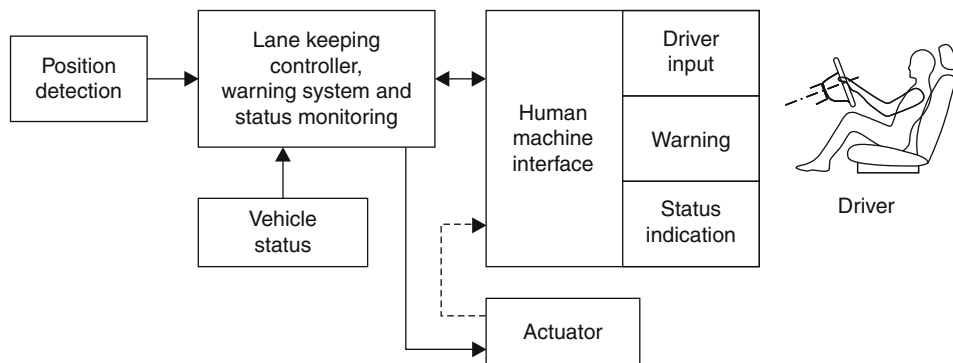
The center diagram of Fig. 26.3 shows a setting of a “tight guidance.” Such a system will apply a considerable assist torque even at small deviations from the lane center. Therefore, the driver will be forced to stay in the center of the lane. Depending on the force of the assist torque an automated lane keeping can be realized with this characteristic.

The right diagram shows a compromise of loose and tight guidance and enables a comfort-oriented control. The driver can wander a little bit within the lane but will receive a perceivable support nearby the lane border. This characteristic is realized by Nissan (introduced in Japan in 2001) and Honda (2002 in Japan, 2006 in Europe). The concept of Honda LKA systems is based on a cooperative operation between driver and vehicle intended to lighten the operation load but at the same time not to diminish driver motivation (Ishida et al. 2003).

Regardless of the torque characteristic most available LKA systems feature a LDW using optical and acoustical warning as a subfunction.

The functional elements of a LKAS are shown in Fig. 26.4. There are additional elements compared to LDW systems but also additional requirements for the functional elements required for LDW.

The most visible differences between LKAS and LDWS functional architecture are the lane keeping controller and the actuator for heading control. The lane keeping controller calculates an appropriate actuator output based on the lateral position and/or the heading of the vehicle within the lane. The curvature of the road segment ahead may also be used as input for the lane keeping controller to realize a predictive or preview control strategy. Because of these considerations typically there are additional requirements for position detection compared to LDWS. All currently available LKA systems use a camera with image processing for lane marker recognition and require visible lane markers on both sides of the lane to fulfill these requirements.



■ Fig. 26.4

Functional elements of a lane keeping assist system

The actuator converts the controller output into a corrective steering to support the driver. Typically a steering torque is applied using the electric power steering or a separate electronically controlled actuator. The assist steering torque can be sensed by the driver giving a tactile feedback on the system action as indicated by the dashed arrow. The detailed characteristic of the lane keeping controller has to be tuned according to the vehicle steering and actuator dynamics in order to achieve stability and comfort requirements.

In addition to a steering torque actuator a selective braking of the wheels as introduced by Nissan for their “*Lane Departure Prevention Systems*” in 2007 and Daimler in 2010 is possible. Further possible actuator principles include superposition steering and adaptation of the suspension characteristic. For these actuator types no direct tactile feedback of the actuator to the driver is given unless it is realized using a separate actuator. Generally speaking, in contrast to LDW systems here a direct access to the vehicle control is given which needs to be considered during the design process as described in [Sect. 3](#).

Additional extensions of LKAS compared to LDWS may include a driver hands-off detection to avoid that drivers take their hands from the steering wheel. In addition driver’s actions might be monitored to detect intended avoidance steering maneuvers (as quick and forceful steering action) and temporarily deactivate LKAS during braking maneuvers.


In the United States, manufacturers seem to hesitate to apply an assist torque to the steering wheel whereas Japanese manufacturers see LKAS as a feature for improved comfort and safety. For European manufacturers the safety aspect seems to be more relevant than comfort improvements. These regional differences might be caused (at least to some extent) by different weighting of steer feel perception in different regions.

2.3 Combined and Integrated Lateral Support Systems

In addition to the pure specification of LDW and LKA as described above there are several combinations possible. For example, Daimler has recently introduced for the European

market a combination of a LDWS with a selective braking of a rear wheel for heading control in case of imminent danger. Imminent danger is defined by passing a solid line or risk of collision with a vehicle on the neighboring lane as detected by radar sensors for Blind Spot Warning (BSW). For passing dashed lines and no risk of collision with another vehicle a warning by a vibrating steering wheel is issued.

3 Development of LDW and LKA Systems

This section addresses the realization of LDW and LKA systems and focuses on aspects to be considered during the development as well as current technological aspects of the used components. The block diagram shown in  Fig. 26.3 is used as reference for the ease of explanation. However, in real systems the mapping of the functional modules to different control units and modules may vary depending on the vehicle architecture.

A systematic development process should be applied considering the current legal framework as well as the published state of the art. First of all regulations need to be considered for vehicle type approval. Standards are not legally binding but describe the current state of the art. Developing a system in accordance with a standard helps to prove compliance with product safety regulation.

For LKAS depending on the actuator there are different regulations applicable. For example, for steering actuators UN ECE R79 Annex 6 defines “*special requirements to be applied to the safety aspects of complex electronic vehicle control systems*” which includes requirements for risk and safety analyses needs to be fulfilled. This regulation is valid for most regions worldwide. In Japan, there is a Technical Guideline on LKAS by the ministry of Land Infrastructure and Transport (MLIT). This guideline has no regulatory force but should be considered for type approval. It specifies some limitations for the speed range, the assistance torque and needs for hands-off detection.

At UN ECE WP29 GRRF an “Informal group on Automatic Emergency Braking and Lane Departure Warning Systems (AEBS/LDW)” has been established. Currently this group discusses requirements for trucks and buses but not for passenger cars. It is expected that the results of this working group will be adopted into several national regulations. For Europe mandatory equipment for trucks and buses is expected from 2015.

For LDWS the ISO 17361 standard titled “*Intelligent transport systems – Lane departure warning systems – Performance requirements and test procedures*” has been published and for LKAS the ISO 11270 is under preparation.

ISO 26262 is a Functional Safety standard which is currently under development, titled “*Road vehicles – Functional safety*.” ISO 26262 is currently (July 2011) in the “*International Standard under publication*” state. Final publication is expected soon. Therefore, most companies developing ADAS are currently preparing themselves for the application of this standard. Within the functional safety process described by ISO 26262 the system status monitoring and risk assessment are of central importance. The *controllability* concept has been introduced in order to consider the human capabilities and limitations for risk assessment.

Within the European Commission funded project “RESPONSE 3” a “Code of Practice” has been developed for the assessment of *controllability* of an advanced driver assistance system (Knapp et al. 2009). The key idea of controllability for LKAS is to assess the likelihood that the driver can cope with LKAS-assisted driving, system limits and system failures.

For example, the question if even the weakest driver can override an LKAS assist torque to initiate an emergency avoidance steering maneuver needs to be considered. Driver’s steering input might be analyzed for quick and forceful steering operation which could be interpreted as an intended maneuver and in consequence LKAS support is temporarily deactivated to avoid interference with driver’s maneuver. Another scenario for a controllability assessment is an unexpected deactivation of LKAS (maybe due to a non-detected lane marker) while driving in a curve. Could a sudden reduction of the assist torque induce instability to the vehicle which cannot be handled by the driver? For example, Volkswagen published extensive studies for their Lane Assist system addressing this aspect (Switkes et al. 2007).

For product liability reasons foreseeable misuse needs to be anticipated and avoided. If, for example, a lane keeping assist system is tuned to enable automated lane keeping this might be (mis-)used by drivers. In such a case drivers might be motivated to read magazines, to take a nap, or to do other attractive things while driving somewhere on a highway in the prairies. In this context, excessive driver trust in a system should be considered and avoided as well because driver’s attention may decrease in a subconscious process as well which will deteriorate his capability to handle a critical situation if required.

Test procedures and protocols of authorities, user organizations, and magazines could have a huge influence on customer perception of a dedicated system, especially compared to competitor systems. So achieving a certain score in a test procedure might be an important requirement for product management. A brief overview on activities of assessment programs is given in [Sect. 4](#).

3.1 Lane Marker Recognition

For today commercially available systems the lane boundaries are represented by visible lane markers. So a reliable detection of the lane markers is the key element for LDW and LKA systems. First lane marker recognition algorithms have been presented decades ago and nowadays there are many different implementations available.

First of all the desired output parameters should be specified depending on the function to implement. Potential output parameters include

- Lateral distance to the lane boundary – *DLC* (Distance to Line Crossing)
- Orientation or heading angle within the lane – ψ
- Rate of departure – V_Y
- Curvature at a given distance in front of the vehicle
- Change rate of curvature at a given distance in front of the vehicle
- Lane keeping performance

Curvature and change rate of curvature might be given as clothoid parameters or as a polynomial approximation. The latter parameters are more relevant for LKAS whereas LDWS often rely on the first parameters. For all parameters the quality in terms of accuracy, availability, and reliability needs to be considered. For LKAS availability of lane marker recognition and thus of the LKA function is an important customer visible property. Availability rates of 90–95% should be targeted for LKA support. It should be noted that availability of lane marker recognition is not exactly the same as availability of LKA function because the LKA controller might filter the signal in the time domain to avoid frequent activation and deactivation. For this purpose statistical filter algorithms like Kalman filters or particle filters are often used.

The lane recognition needs to be designed to recognize the lane markers of the target markets for a specific vehicle. Lane markers vary from country to country significantly. For example, the lane marker width varies from 80 to 300 mm. There are single and double solid and segmented lines. Further parameters are the length of the segments (for interrupted lines) and the gaps (voids) between the segments. In several states of the United States so-called *Botts' dots* (also called “bot dots”) are used. These are raised pavement markers with a diameter of about 10 cm and rectangular reflectors. In addition you can find white, yellow, blue, and red painted lane markers. All these variations encode a different meaning in the particular country. Depending on the function concept the LDW or LKA system needs to detect and interpret these meanings. For example, in some regions temporary markings in construction zones are encoded by a defined color. If a system should provide support in construction zones color differences need to be detected.

It should be noted that painted strips inherently give information on the lane direction (in contrast to *Botts' dots*). This property could be used for a lane recognition algorithm, for example, to estimate the curvature ahead.

Furthermore ranges of geometric parameters of the lanes and the covered vehicle dynamic need to be defined. This includes minimum and maximum values for the lane width, the curvature, and banking. For the vehicle dynamic the range of heading angle ψ , moving direction v , and lane departure rate need to be specified – at least implicitly. The vehicle velocity needs to be specified for functional aspects (for which speed range a support is given) as well as for lane recognition specification. First of all the vehicle moving direction v and the lane departure rate are coupled via the vehicle longitudinal velocity and in addition the vehicle velocity may determine computational requirements for a tracking of image processing features if applied. In addition increased velocity will increase motion blur and thus determine requirements for exposure time as well as frame rate.

In addition to the previously described parameters the lane keeping performance can be extracted from lane marker recognition. This parameter might be used to adapt warning threshold in order to reduce false alarms but might also be used as an input for drowsiness detection or driver attention monitoring (McCall and Trivedi 2006).

The environment and lighting conditions for lane marker recognition may vary in a wide range. Direct sunlight, nighttime, tunnels, rainy weather, and further conditions

need to be considered. Depending on the road surface condition there might be disturbing reflections. Under some conditions it might happen that a white lane marker appears darker than the pavement. Further properties include photometric properties of new and worn-out lane markers and the non-homogeneity of the road surface in the target market.

Whereas early approaches for vision-based lane recognition are based on detecting and tracking the lane markers recent approaches aim to “understand” the scene ahead. Information on vehicles traveling ahead give information on areas where lane markers are occluded by these vehicles (McCall and Trivedi 2006).

Currently there are two technical approaches available on the market: camera-based detection and infra-red laser-based detection.

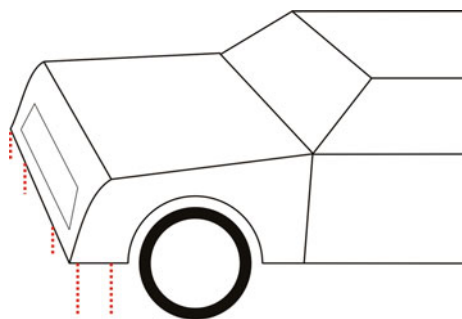
The latter approach is realized by supplier Valeo for a lane departure warning system for several Citroën and Peugeot models in Europe. In total six sensors are mounted in the front bumper to scan the road surface underneath the front of the vehicle (● Fig. 26.5).

Because of the shown sensor configuration this system can sense the actual road markers at the vehicle location but not sense the road geometry ahead. Therefore, such a sensor configuration can be used for LDWS but not for LKAS. Furthermore the LDW criterion is purely based on the lateral distance to a lane marker. A time to line crossing criterion cannot be realized with this sensor configuration. Furthermore this sensor configuration is not applicable for roads with Botts’ dots because you might leave a lane without “seeing” a dot with these sensors.

By mid of 2010 the majority of LDW and LKA systems use a single monochrome camera mounted nearby the inner rear view mirror as sensor for lane marker recognition. There are first systems using a color camera and in future color cameras will be used more widely. Further parameters are sensitivity (especially for night-time driving), dynamic range, and spatial resolution.

For more information on image processing algorithms for lane marker recognition please refer to (McCall and Trivedi 2006; Gern et al. 2002; Smuda von Trzebiatowski et al. 2004; Dickmanns 2007; Franke et al. 2001).

The described lane marker recognition methods described so far rely on visible lane markers. In addition to this there are approaches using other types of sensors and/or



■ Fig. 26.5
Sensor configuration using infra-red laser sensors

infrastructure, for example, in (Thorpe et al. 1997) approaches for lateral control by magnetic nails or buried cables carrying an audio-frequency signal. These approaches require an additional dedicated infrastructure in addition to the visible lane markers initially applied for human drivers. As an advantage these systems can work even under poor visibility or covered roads. Therefore, the magnetic nail approach has been examined for snow plows (Tan and Bougler 2001). However, these systems are not yet widely introduced and even a wide introduction cannot be expected for the future.

Other approaches use signals from radar or laser range sensors to acquire information on the path ahead. Typically these signals are used as prior information for vision-based algorithms. Furthermore, the use of GPS and navigation map data is possible as prior information for a vision-based system. The exclusive use of GPS data is currently and for the next years limited by position and map accuracies.

3.2 Warning Elements

The warning element is the most important element of the human machine interface of a LDW system. The primary purpose of the warning element is to alert the driver of an imminent lane change and urge him to take corrective actions. First of all a warning element needs to reach driver’s attention through one of the human senses. For LDW systems the visual, auditory, and haptic (or tactile) channels are used. The olfactory sense and the sense of taste are not used for advanced driver assistance systems even if this might be a nice imagination.

In addition to attracting driver’s attention a warning element should preferably indicate the nature of the danger and if possible indicate the direction where the danger comes from. For example, a vibrating gas pedal might raise driver’s attention but will not indicate a problem with lateral control of the vehicle. ♦ Table 26.1 lists warning elements currently used for LDW and some of their properties. A combination of several elements is possible and also realized especially for display indication and beep sounds.

■ Table 26.1
Examples of warning elements used for LDWS (column “Type” indicates if a directed (D) or undirected (U) warning is issued by a specific element)

Warning element	Addressed sense	Type	Used by
Steering wheel vibration	Tactile	U	BMW, Daimler
Vibration under left and right seat cushion	Tactile	D	Citroën, Peugeot
Stereo sound imitating a rumble strip	Auditory	D	Mercedes trucks (United States: Freightliner)
Beep sound	Auditory	U	Honda, Toyota
Display indication	Visual	U	Honda, Toyota
Steering torque	Tactile	D	Honda, Toyota, VW

A directed warning indicates the direction of the lane departure which enables a quicker reaction at least subjectively. The directed steering torque provides both a steering assist for LKA function as well as tactile information. Depending on the characteristic of the torque this information might be perceived as a warning. Especially for systems featuring a torque characteristic as shown in ► Fig. 26.2 like the Volkswagen Passat CC the assist torque will give a warning with recommendation of action.

According to ISO 17361 “an easily perceivable haptic and/or audible warning shall be provided.” Furthermore it is recommended (as “optional function”) that “if the haptic and/or audible warnings are not designed to indicate the direction, a visual cue may be used to supplement the warning.” Therefore, many available LDW systems use a combination of multiple warning elements to realize a *warning strategy*. Visual warnings are typically shown in the instrument display or in a head-up display if equipped.

Finally the selection and tuning of the used warning elements will influence the *effectiveness* of a warning strategy in case of a true positive warning. The minimum effectiveness is an important requirement which needs to be fulfilled by a selected warning element. Effectiveness can be measured in terms of success rates for avoiding unintended

■ Table 26.2

Selected requirements for warning elements

Requirement	Description/remark
Effectiveness	Success rates for avoiding unintended lane departures
Reaction time (distribution)	Time for driver's steering reaction after a warning has been issued
Physical integrity	A warning element needs to be designed in a way to avoid any harm to the driver or other occupants. For example, the South African vuvuzelas are not recommended as auditory warning element because they might cause a hearing disorder (Swanepoel et al. 2010)
Exposure/unmasking of the driver	If a warning element can be perceived from passengers of a vehicle driver's driving skill might be unmasked. This effect needs to be considered for justified as well as for unjustified warnings
Uniqueness and type of warning	The design of the warning element should intuitively inform the driver on the urgency of the event to be warned of. According to ISO 17361 “the warning shall be clearly distinguishable to the driver by a haptic, audible, or visual modality, or any combination thereof.” For example, the warning about an imminent lane departure should not be the same as the indication of empty washer fluid
Subjective effectiveness	The subjective effectiveness indicates the confidence of drivers in a system to contribute to road safety. It can be estimated by questionnaires whereas the (objective) effectiveness is determined by analyzing driver's behavior
Objective disturbance	Depending on the design of the used warning element there might be knee-jerk driver reactions. This might provoke dangerous situations in case of unjustified warning and thus needs to be minimized

lane departures or to avoid accidents caused by an unintended lane departure. Such an assessment could be quite tricky and expensive as it requires lengthy subject studies with different driver types under different traffic, road, and driver conditions. The *reaction time (distribution)* is another property of the effectiveness which might be assessed easier. “Just” you have to define how fast is fast enough.

Besides the effectiveness there are numerous further properties and corresponding requirements. These requirements may originate from aspects like acceptance and product liability. If, for example, a warning element is designed in a way to cause panic-like reactions there might be an accident because of a false alarm. This needs to be avoided.

❖ **Table 26.2** summarizes some requirements for warning elements.

Some of the described requirements are contradicting. So it is an important task to define the weighting of the requirements according to the targeted vehicle type and brand image. For example, exposure of the driver will probably not be accepted for a bus application (because it might deteriorate the confidence of passengers in driver’s driving skills) or a sporty vehicle (due to the ego of the driver). Furthermore, regional and brand image differences may influence the weighting. For example, beeps are common for some Japanese brands but German premium manufacturers aim to minimize them.

3.3 Lane Keeping Assistance Controller and Actuator (Just LKAS)

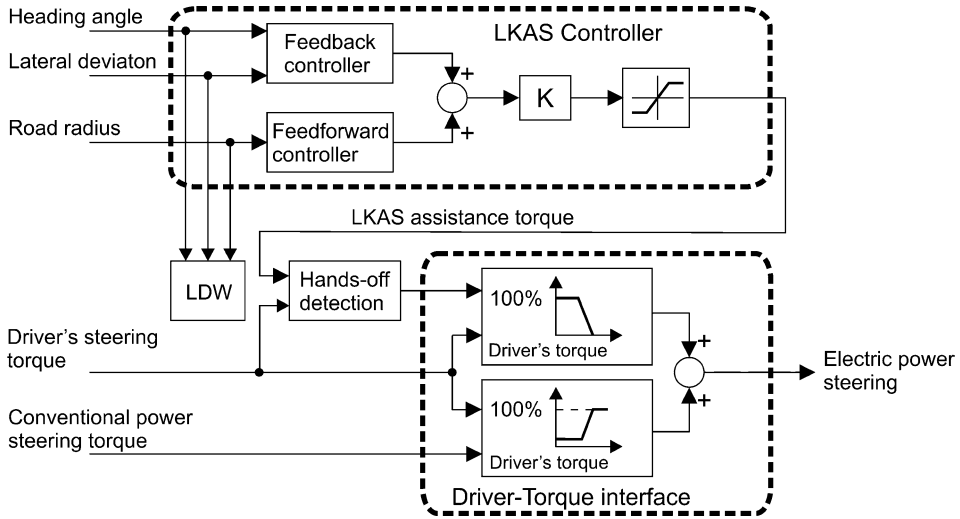
Main requirements for the LKAS controller include

- Providing effective support to the driver for keeping the lane
- No patronizing the driver
- Enabling overriding by the driver at any time
- Ensuring stability in all situations (including avoidance of saw-tooth behavior at banked roads)
- Avoiding excessive driver dependence on the system

Some of these requirements are mutually contradicting as, for example, preventing excess driver dependence on the system while at the same time enabling maximum advantage of the assistance effect.

❖ **Figure 26.6** shows the control architecture of the Honda LKA system as example. Main inputs for the controller are the outputs of lane marker recognition module. The shown controller uses lateral offset and orientation within the lane as well as the road curvature as inputs for combined feedback and feed-forward control to determine the steering torque required to keep the vehicle in the lane. In order to keep the driver maintaining a constant steering operation just a ratio K of the determined steering torque is applied as assistance torque. A ratio of 80% is for example used by Honda (Ishida et al. 2003). In addition a limiter operation is applied to ensure possibility of overriding by the driver.

The calculated assistance torque is fed through the hands-off detection and the driver torque interface to the steering controller which converts the controller output into a lane



■ Fig. 26.6
Block diagram of Honda LKA controller

keeping assist torque. As actuator an electric power steering system is used. For larger vehicles a separate electronically controlled actuator in combination with a hydraulic power steering is possible.

For hands-off detection steering driver's movements are analyzed and within 5–10 s a hypothesis whether the driver is steering is tested. The driver torque interface detects intentional driver maneuvers and temporarily suspends LKA support to avoid interference with driver's operation.

For Lane Departure Prevention Systems using the rear wheel brakes (as introduced by Nissan and Daimler) the block diagram of the controller will be different compared to Fig. 26.6 but most of the considerations remain valid.

4 Market Trend and Outlook

Currently LKAS and LDWS are offered for many mid- and upper class passenger cars but have not yet reached the popularity of leather seats and alloy wheels even if the latter features often are more expensive. Most manufacturers applying these systems achieve just single digit percentage equipment rates for these options. Key activities for increasing the penetration rate of ADAS include

- Regulation
- Cost reduction and/or increased functionality
- Increased awareness
- Assessment of safety benefit – and appropriate ratings
- Incentives by government or insurance premiums

Mandatory equipment of LDWS as expected for trucks and buses in Europe will stimulate the market and enable economies of scale. Also the passenger car market might benefit from this. Regulatory equipment of passenger cars is currently not foreseeable for the next 5 years.

Cost reduction was a major challenge for the first introductions of LDWS and LKAS. In the meantime big progress was made for cost of electronic control units (due to falling prices for semiconductors). Nevertheless further decrease of total system costs is required including cost of cabling and interfaces. Here standardization of components and interfaces could pave the way for cost reductions. For LKA systems the integration in the chassis control should be simplified to enable an easy and safe integration in many vehicle segments.

To raise customers' awareness of LDW and LKA systems is another prerequisite for increasing the penetration rate of these systems. About 10 years after the first market introduction still many potential customers do not know about these systems and their availability even for mid-size passenger cars. All communication channels will be used including product advertisement but also press and internet publications, information by user organizations, and more. For example in Europe, "awareness raising actions" are on the agenda of the European Commission and a TV documentation is sponsored (Bishop 2011).

Hearing and knowing about a certain system is one side, and being convinced on the benefit and finally buying a system is another side. The various "New Car Assessment Programs" (NCAP) in different regions worldwide had a big influence on the introduction of passive (or secondary) safety systems. In recent years, NCAP organizations in the United States and Europe have been active to set up rating schemes for active (or primary) safety systems like LDWS. It can be expected that NCAP ratings of LDWS will raise user awareness and hence popularity of these systems.

The US National Highway Traffic Safety Administration (NHTSA) recently published a test procedure titled "Lane Departure Warning System NCAP Confirmation Test" under its NCAP assessment program. This document describes test procedures for painted lane markers as well as the raised pavement markers called "Botts' dots" (also called "bot dots").

Euro NCAP has introduced the "Beyond NCAP" assessment method in early 2010. The Beyond NCAP procedure describes requirements for a dossier to be submitted by a vehicle manufacturer for a specific vehicle equipped with a specific ADAS. In such a dossier accident statistics showing the potential safety benefit, detailed accident mechanism including driver behavior and finally a calculation of expected benefit on vehicle level needs to be shown. The dossier will be evaluated by an assessment group within 6 months. This procedure targets to develop a test procedure with evidence for the expected safety benefit in real-world-driving conditions. Based on the "beyond NCAP" method definition of a Euro NCAP test protocol for LDWS and/or LKAS can be expected within a few years.

In most regions worldwide unintended lane departures are a major cause of severe accidents. Some publications suggest ratios between 25% and 50% of all severe accidents. However, as described in the introduction the root causes and accident mechanisms of

unintended lane departures might vary significantly. A certain LDWS or LKAS will target just a subset of all accidents-related lane departures. For example, today's LDWS cannot recognize lane boundaries in urban environments and hence will not support avoiding accidents in urban areas as well as in tight curves. Generally LDWS cannot override the laws of physics and thus will not help in case of excessive speed in a curve.

All these aspects on assessment of LDWS and LKAS mentioned above are closely linked to the question of the effectiveness in reducing accidents. The assessment of advanced driver assistance systems poses many methodological challenges compared to the assessment of passive safety systems. Passive or secondary safety system like airbags or a defined crumple zone are typically not perceived by the driver while driving and thus an influence on the driving style and the accident configuration can be neglected. In consequence the effectiveness of these systems can be assessed using crash test dummies in a crash facility. On the other hand driver assistance systems like LDW and LKA directly intervene in the driving process and interact with the driver. Therefore, an influence on the driving style has to be expected but is difficult to quantify. Possible approaches for the assessment include theoretical analyses, tests on proving ground, field operational tests, and analysis of statistical real-world data. All these methods have advantages and disadvantages and probably, finally a combination of these approaches is required to give reliable results. In (Page et al. 2007) a discussion on different methods and a comprehensive literature overview is given. Especially in the ramp-up phase of new systems, real-world data is rarely available and strongly biased by several effects like the "population effect" caused by a specific population deciding to buy a vehicle with a certain driver assistance system, and the "model year effect" describing the fact that new vehicle models typically feature several safety improvements and positive safety effects cannot be assigned to a single system easily.

Therefore, in the ramp-up phase field operational tests play an important role for the assessment. One of the larger field operational tests for LDWS has been conducted in 2006 in the Netherlands (Alkim et al. 2007).

An in-depth analysis of published research on the safety effect of advanced driver assistance systems was done by the European Commission funded project ASSESS (ASSESS 2010) reporting safety benefits of 13.5% for fatalities and 9.5% for injuries. Even these figures indicate the major safety potential of these systems. If the indicated safety potential will motivate insurers to grant discounts on insurance premiums or national authorities to grant tax incentives is difficult to predict.

Today available LDW and LKA systems are offered as optional feature for several vehicles as listed in the introduction. Lane departure warning systems can be applied relatively easily in a vehicle since they do not require an application to the chassis control (steering and/or braking). Therefore, many manufacturers offer LDW systems. Even aftermarket solutions are available. All current systems are designed for operation on extra-urban highway-like roads or smooth country roads. ➤ Table 26.3 lists typical parameters of the working range of today's systems.

The upper speed range is given as a technically supported velocity. In many regions there is a limitation by regulation. The lower-supported speed range may differ from the lowest activation speed. In order to avoid frequent activation and deactivation a hysteresis

■ Table 26.3
Range of parameters for actual systems (the upper speed range depends strongly on national speed limits)

Parameter	Typical values	
	LDWS	LKAS
Lower-supported speed range	60–70 km/h	60–70 km/h
Upper speed range	Max. vehicle speed	170–180 km/h
Maximum steering assist torque	–	2–3 Nm
Supported lane width	3–4 m	3–4 m
Minimal curve radius	230 m	230 m
Maximum lateral acceleration	–	0.2 G

might be applied. For example, the Volkswagen lane assist activates at 65 km/h but deactivates at 60 km/h.

As described before the majority of today’s lane departure warning and keeping systems use a single monochrome camera for detecting the lane markers of the lane ahead. Furthermore they are restricted to roads with clearly visible lane markers. For the future overcoming several limitations of today’s systems can be expected.

In some markets construction highway zones are realized by narrowing the lanes using yellow lane markers. Sometimes the left lane has a width of just 2 m. For these narrow lanes many drivers would like to have support. A combination of color camera and forward looking radar might be suitable to handle these situations.

Illumination and weather conditions heavily influence the visibility of lane markers in the images captured by a camera. Especially wet road surfaces at nighttime can often not be covered by today’s systems. Progress of camera technology in terms of increased sensitivity, increased dynamic range, and additional color information as well as improved image processing algorithms will help to cover further environment conditions within the next years.

Additional sensor data like radar, laser range sensors, and navigation map data could provide information on the road geometry and thus contribute as well to improved robustness against illumination conditions but also help to release some geometry constraints like the minimum curvature. Finally precise three-dimensional road information as acquired by forward looking radar and laser sensors but also by stereo camera systems may enable road keeping support for road segments without visible lane markers. At the same time camera sensors which have initially been used for lane marker recognition exclusively will be used for further functions including

- Speed limit warning by recognizing speed limit signs (a function popular in Europe)
- Advanced lighting functions like high/low beam assist
- Pedestrian recognition
- Support of *Adaptive Cruise Control* (ACC) by improving lane assignment of detected objects

From a functional point of view a closer integration with other advanced driver assistance systems is expected for the future. The system recently launched by Daimler combines LDW with a *Blind Spot Warning* function (see ► Sect. 2.3) and shows a first stage of extension. A combination with systems like ACC and *Forward Collision Mitigation Brake* is also possible including detection of oncoming vehicles which could improve safety especially on country roads.

About 10 years after introduction of the first LDW and LKA systems many new levels of functionality will enter the market while at the same time end customer prices are decreasing. Thus increasing penetration rates can be expected for the upcoming years contributing to safer and more comfortable road traffic.

5 Conclusion

Assistance systems for supporting the driver in keeping the lane have been subject of research since several decades. The driver support could be achieved by a warning or an active support in keeping the lane, for example, by an assistance steering torque. Since about 10 years the first systems are commercially available in slightly different functional specifications. The basic concepts and the characteristics of these functional specifications have been described as well as their implications on the required functional elements. For selected functional elements like the lane recognition possible technical implementations are briefly described.

An overview on requirements for the development of lateral support systems is given. This includes references to regulations and standards for functional characteristics and the development and validation process. A discussion on contradicting requirements for warning elements gives an insight in the complexity of the actual product development. Finally a market overview and outlook are given.

References

- Alkim T, Bootsma G, Looman P (2007) The assisted driver – systems that support driving. Publication of the Dutch Ministry of Transport, Public Works and Water Management, Rijkswaterstaat. http://www.fot-net.eu/download/seminars/LONDON051010/the_assisted_driver.pdf. Accessed 17 July 2011
- ASESS (2010) Assessment of integrated vehicle safety systems for improved vehicle safety. EU Project deliverable. <http://www.assess-project.eu/site/en/documents.php>. Accessed 24 June 2011
- Bishop R (2011) Smartest Cars Video Project. Final project report. http://ec.europa.eu/information_society/activities/esafety/doc/rtd_projects/fp7/scvp_final_report.pdf. Accessed 24 June 2011
- Broggi A, Bertozzi M, Conte G, Fascioli A (2001) ARGO prototype vehicle. In: Vlacic L, Parent M, Harashima F (eds) *Intelligent vehicle technologies*. Butterworth Heinemann, Oxford, pp 445–493
- Dickmanns ED (2007) *Dynamic vision for perception and control of motion*. Springer, London
- Franke U, Gavrila D, Gern A, Goerzig S, Janssen R, Paetzold F, Woehler C (2001) From door to door – principles and applications of computer vision for driver assistant systems. In: Vlacic L, Parent M, Harashima F (eds) *Intelligent vehicle technologies*. Butterworth Heinemann, Oxford, pp 131–188
- Gern A, Moebus R, Franke U (2002) Vision-based lane recognition under adverse weather conditions using optical flow. In: 2002

- IEEE intelligent vehicles symposium, Versailles, France
- Ishida S, Tanaka J, Kondo S, Shingyoji M (2003) Development of a driver assistance system. In: 2003 SAE world congress and exhibition, Detroit
- Knapp A, Neumann M, Brockmann M, Walz R, Winkle T (2009) Code of practice for the design and evaluation of ADAS. Final project report, ACEA, Belgium. http://www.acea.be/index.php/news/news_detail/acea_endorses_response_code_of_practice_for_advanced_driver_assistance_syst/. Accessed 24 June 2011
- Mammar S, Glaser S, Netto M (2006) Time to line crossing for lane departure avoidance: a theoretical study and an experimental setting. *IEEE Trans Intell Transport Syst* 7(2):226–241
- McCall JC, Trivedi M (2006) Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation. *IEEE Trans Intell Transport Syst* 7(1):20–37
- Page Y, Rivière C, Cuny S, Zangmeister T (2007) A posteriori evaluation of Safety Functions effectiveness – Methodologies. EU Project TRACE deliverable. <http://www.trace-project.org/publication/archives/trace-wp4-d4-2-1.pdf>. Accessed 17 July 2011
- Pomerleau D (1995) RALPH: rapidly adapting lateral position handler. In: 1995 IEEE symposium of intelligent vehicles, Detroit, pp 506–511
- Smuda von Trzebiatowski M, Gern A, Franke U, Kaeppler U, Levi P (2004) Detecting reflection posts – lane recognition on country roads. In: 2004 IEEE intelligent vehicles symposium, Parma, Italy
- Swanepoel DW, Hall J, Koekemoer D (2010) Vuvuzela – good for your team, bad for your ears. *S Afr Med J* 100(2):99–100
- Switkes J, Gerdes JC, Schmidt G, Kiss M (2007) Driver response to steering torque disturbances: a user study on assisted lanekeeping. *Advances in Automotive Control*, vol 5. Elsevier, Seascope Resort, USA
- Tan HS, Bougler B (2001) Vehicle lateral warning, guidance and control based on magnetic markers. PATH report of AHSRA smart cruise 21 proving tests, California PATH working paper, UCB-ITS-PWP-2001-6
- Thorpe CE, Jochem T, Pomerleau D (1997) Automated highways and the free agent demonstration. In: International symposium on robotics research. Springer, Hayama, Japan, pp 246–254
- United Nations (1968) Convention on road traffic. UN ECE, Geneva
- Van Winsum W, Brookhuis KA, De Waard D (2000) A comparison of different ways to approximate time-to-line crossing (TLC) during car driving. *Accid Anal Prev* 32:47–56. Elsevier

27 Integral Safety

Klaus Kompass¹ · Christian Domsch¹ · Ronald E. Kates²

¹BMW Group, Munich, Germany

²REK-Consulting, Otterfing, Germany

1	<i>Introduction</i>	710
2	<i>Limitations of Passive Safety Systems</i>	713
3	<i>The Approach of Integral Safety</i>	715
4	<i>Quantification of Field Effectiveness</i>	721
5	<i>Summary and Conclusion</i>	726

Abstract: In developed countries such as the USA or Europe, the risks of injury or fatality in traffic accidents have declined significantly in recent years. These reductions apply to both vehicle passengers and other involved persons. Much of this improvement has been attributable to progress in the field of *passive safety*, i.e., better protection of car occupants in situations where an accident is unavoidable. However, the marginal benefits resulting from additional efforts and expenditures in passive safety have begun to decrease; in other words, a classical “point of diminishing returns” has been reached. Increasing emphasis for achieving further significant improvements in vehicle safety will be placed on *integral* safety systems: Integral safety involves a concerted strategy of interlinking sensors and actuators of active and passive safety. The primary goal of this interlinking is optimization of performance and robustness of safety systems for occupants, but integral safety approaches can also achieve better protection of vulnerable road users than passive safety measures alone. In view of considerations such as reduction of CO₂ and fuel consumption, there is another attractive benefit: integral safety can serve to reduce the steady weight increase of vehicles and thus provide an important contribution to the development of both sustainable *and* safe vehicles.

In order to develop effective measures for mitigating the severity of traffic accidents or even completely avoiding them, it is essential to understand the mechanisms of accident events, including the processes and risks involved in traffic situations in which these accidents occur. A quantitative understanding of these processes and risks aids in assessing the potential effectiveness of vehicle safety measures. The automobile industry is faced with enormous challenges in discovering and implementing the most effective solutions. Assessment by legal authorities and/or consumer groups should concentrate on safety performance, not on specification of particular technologies or methodologies, and should encourage implementation of devices providing greatest safety benefits by mandating robust and standardized testing and assessment techniques that quantify and measure effectiveness independently of technological details.

1 Introduction

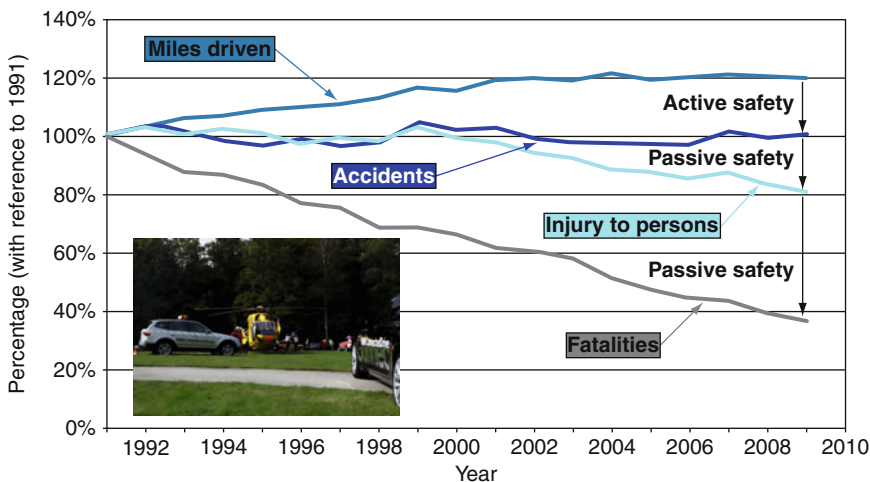
The term “Vision Zero” (Vision Zero Initiative 2011) today is a synonym for future targets in traffic safety. The vision is based on the idea of avoiding accidents completely, that is, zero accidents with zero injuries and zero fatalities. However, this goal cannot be achieved by a single technological leap, but requires a combination of strategies based on a profound understanding of accident processes.

The generally accepted definition of passive safety includes all features and functions designed to protect victims of traffic accidents during and after the point of no return, that is, when the accident can no longer be avoided. Active safety on the other hand describes features and functions with the primary purpose of preventing such accidents or at least mitigating their severity. Integral safety is a rather new terminology that describes the combination of both active and passive safety and thus provides a cross-link between the

situations before, during and after a possibly injurious collision. Below, this cross-link will be described in further detail.

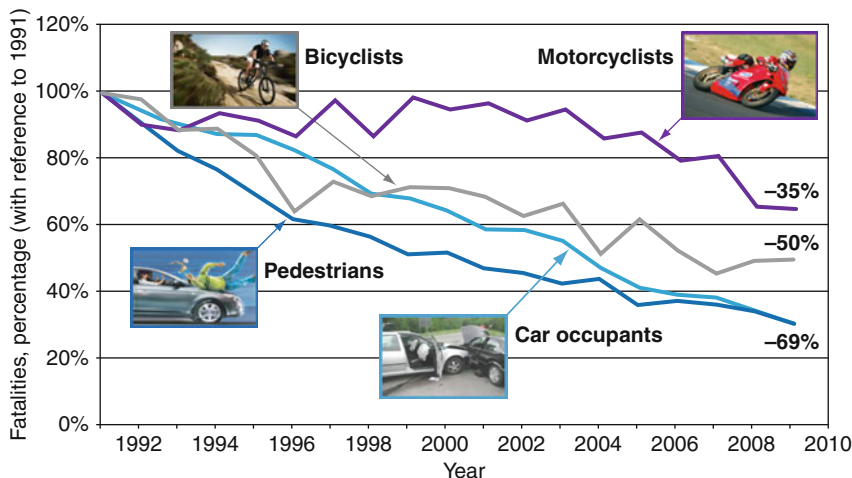
Trends in accident statistics indicate substantial progress in road safety during recent years in the USA, Japan, and the developed countries of Europe. The picture is more complicated in emerging countries such as India or China, which are outside the focus of this report: in these countries, rapidly growing motorization – coupled with a rather immature traffic infrastructure – produces a considerably different situation. A large proportion of progress in “Western” traffic safety has been attributable to impressive developments in vehicle safety, especially in passive safety. Since the introduction of the three-point belt in 1960 – the most effective safety feature at that time in cars – the development of vehicle safety has been characterized by a sequence of milestones: safe passenger cells with energy absorbing crumple zones in the front end, air bags, child seats with standardized attachments, seatbelt pretensioners, and load limiters. Passenger protection by passive safety systems has attained a very high level during the past 30 years, with significant reductions in injuries, especially fatal injuries. With additional penetration of the market by modern vehicles, further improvements can be expected in the near future.

Even though the miles driven have gone up significantly since the 1980s, the number of accidents has remained nearly steady during the last 20 years. The major reasons for this positive fact are improvements in traffic infrastructure, better education of the drivers, and last but not least better active safety performance of the vehicles, that is, better brakes, better suspension, vehicle stability systems, etc. (● Fig. 27.1).



■ Fig. 27.1

Influence of active and passive safety – statistical trends in German accident frequencies (DESTATIS 2010)



■ Fig. 27.2

Development of fatalities – differences depending on categories of road users (DESTATIS 2010)

However, not all road users benefit equally from higher traffic safety (► Fig. 27.2). Car occupants are highly protected by passive safety features. Due to this fact, the number of fatalities has decreased significantly.

Interestingly, the decrease in pedestrian fatalities has been comparable to the decrease in fatalities of car occupants, even though dedicated protection systems for pedestrians had low penetration during this period. Possible contributing factors in reduced pedestrian accident severity include better brakes and brake assistance systems. Note that the slope of the curve for pedestrians appears to have been less steep in recent years than for car occupants. Bicyclists and motorcyclists have registered weaker declines in fatalities; further efforts of vehicle manufacturers, infrastructure providers, as well as better training and supervision are essential for greater improvement.

To continue these generally positive trends into the future, targeted measures in vehicle safety will be required. An intelligent combination of active and passive safety elements will contribute to further reductions of accidents, injuries, and fatalities.


Current ADAS (Advanced Driver Assistance Systems) offer the opportunity to evaluate the traffic environment with increasing reliability. Sensor-based systems detect and classify objects, track attributes such as distance and dynamic data, identify high-risk, critical driving situations and initiate suitable responses, such as warning the driver if there is sufficient time for him to avoid an accident by an appropriate controlled maneuver. (In pedestrian conflicts, a few precious tenths of a second delay in impact can suffice to allow a pedestrian to get out of harm's way.) If the traffic situation becomes even more critical and the driver has not yet initiated an intervention, automatic braking can reduce the impact speed and thus the kinetic energy of the collision.


In addition to these measures, optimization of the postcrash phase has a strong potential for achieving further benefits: Medical studies of emergency and hospital care

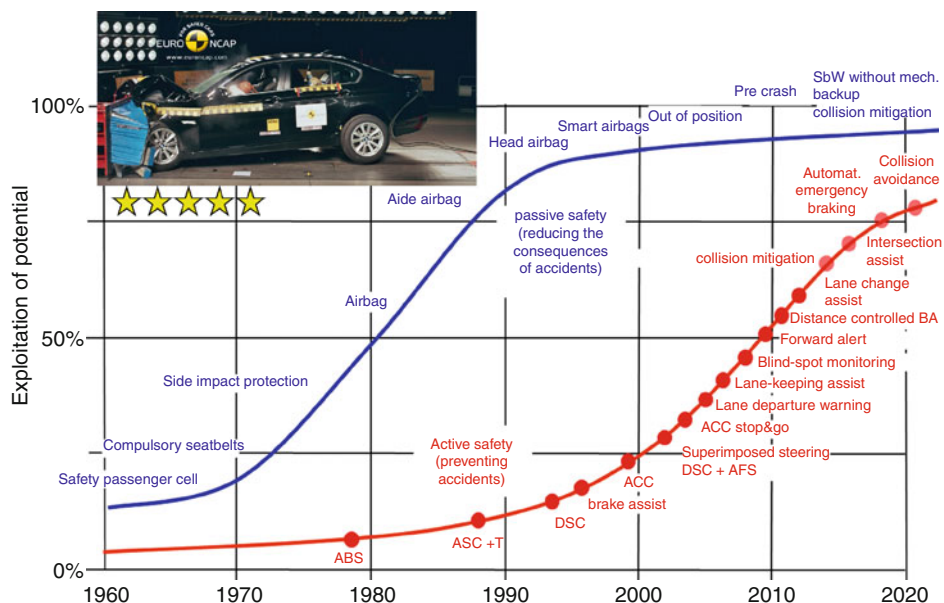
following an accident have shown that rapid rescue and intense, accident-specific care of the most severely injured persons in special trauma centers can significantly increase the survivability of severe car crashes. A prerequisite for delivering rapid, optimal emergency care is to identify risk of severe injury during (and immediately following) the crash using vehicle sensors and to communicate this risk to the call center automatically by an eCall. Thus, a holistic approach will be taken in the next generation of safety systems, integrating different systems and taking all phases of the accident into account: the precrash, the in-crash, as well as the postcrash phases. Today a one-sided concentration on particular aspects of vehicle safety is no longer adequate for achieving further significant improvements. An integral safety approach combining and cross-linking all accident phases promises greater benefits.

2 Limitations of Passive Safety Systems

During the last few decades, passive safety systems have been intensively developed. Recently, the Insurance Institute for Highway Safety IIHS in the USA tested the crashworthiness of a 1959 Chevrolet Bel Air and compared it to a 2009 Chevrolet Malibu (IIHS 2011). One of the impressive results was to see the passenger compartment of the Bel Air collapse and trap the occupants, whereas the driver and passenger of the Malibu had a good chance of sustaining only mild injuries in an intact passenger compartment. Today a modern vehicle such as the BMW 5 series (Model year 2010+) satisfies the highest requirements in different test procedures worldwide. In the most prominent assessments, Euro-NCAP (**E**uropean **N**ew **C**ar **A**ssessment **P**rogram) and US-NCAP, this car achieved the top rating of five stars. The load cases in these test procedures nearly completely represent real-life accidents, considering the impact speed of the vehicle.

These tests together with  Fig. 27.3 illustrate the degree to which exploitation of the potential of passive safety has reached a plateau or a “point of diminishing returns”: Even extensive additional efforts in this area would provide only rather small further improvements. Not only do additional passive safety components fail to provide measurable additional injury reduction, they also worsen the vehicle weight ratio and thus are counterproductive for fuel economy and CO₂ reduction.

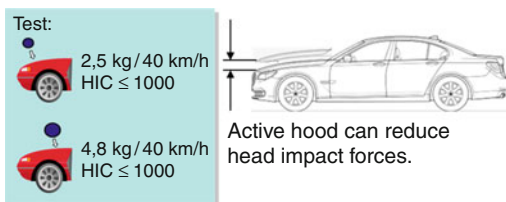
The protection of vulnerable road users such as pedestrians or bicyclists is probably the most difficult challenge in reducing severe injuries and fatalities. Up to now, regulatory as well as NGO safety assessments have defined testing procedures that focus solely on passive safety measures to address situations in which a pedestrian is struck by a vehicle: Human leg impactors or head forms are propelled onto the vehicles surface and accelerations, and forces or bending moments are measured; the car should absorb as much of the impact energy as possible by deformation (Euro-NCAP 2011). Considerable effort has been devoted to energy absorbing construction of the hood leading to elaborate front-end designs. However, in-depth studies have shown that head impact on the hood occurs in <6% of pedestrian accidents ( Fig. 27.4). Moreover, passive pedestrian protection only



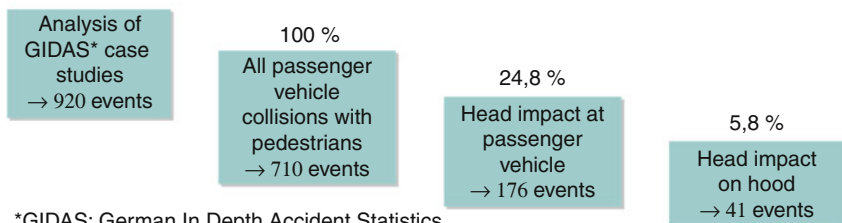
■ Fig. 27.3

Exploitation of potential of active and passive safety systems

Example: head impact on hood:



- Less than 6 % of the head impacts are on the hood
- Passive pedestrian protection only addresses the primary collision
- No influence on secondary collision (e.g. road surface, kerb)



*GIDAS: German In Depth Accident Statistics

■ Fig. 27.4

Pedestrian protection – head impact on hood with only marginal benefit of passive safety systems

addresses the primary collision with the vehicle. Secondary collisions with the road surface or curb occur frequently and cannot be mitigated by this approach.

Summarizing, current passive safety systems offer a high level of vehicle occupant protection in traffic accidents. Advanced restraint systems reduce forces on passengers sufficiently to avoid exceeding biomechanical limits even in severe collisions. At present, restraint systems need to be redesigned for increasingly restrictive legal and regulatory demands (e.g., new FMVSS 214, FMVSS 208 Ph. 2 und Ph.3, Euro NCAP and US NCAP Overall Rating). These demands require increasingly sophisticated adaptation in passenger-restraint systems, which are tested by measurements on dummies in laboratory crash tests. At the same time, it is becoming more and more difficult to demonstrate that these additional efforts are effective in real accidents in the field: it appears that we are approaching a point of diminishing returns in passive safety systems. Of course, improvements in specific details are always possible, but great strides are not to be expected in the isolated development of passive safety systems. New approaches are needed (Kompass and Huber 2009).

3 The Approach of Integral Safety

Functionalities in the area of active safety have considerable additional potential to improve vehicle safety. The increasing application of Advanced Driver Assistance Systems (ADAS) has set the stage for development of active safety systems. The penetration rate for ADAS has been steadily increasing, particularly in higher-valued vehicles, and significant growth in lower-priced vehicles is expected. Some accidents can be entirely prevented by active safety systems. However, even in cases where the accident is unavoidable, the mitigation effect in alerting the driver and thus potentially reducing the impact speed is beneficial per se. In the following, the mechanisms of collision mitigation and collision avoidance will be illustrated based on the example of pedestrian protection.

In the foreseeable future, advanced driver assistance systems will contribute substantially to improvements in traffic safety by collision mitigation and the reduction of crash severity, as well as by warnings and assistance functionalities, which serve to enhance and augment the driver's own capabilities.

The term *integral safety* as used here describes a holistic approach linking the fields of passive and active safety, which up until now have generally been treated as separate subjects. Active safety systems perform prediction and assessment of impending accidents and enable preparation and improved protection of the vehicle and the driver for the collision. In order to reduce the number of accident victims and the severity of injuries substantially during the next 10–20 years, the integral safety approach offers two key strategies.

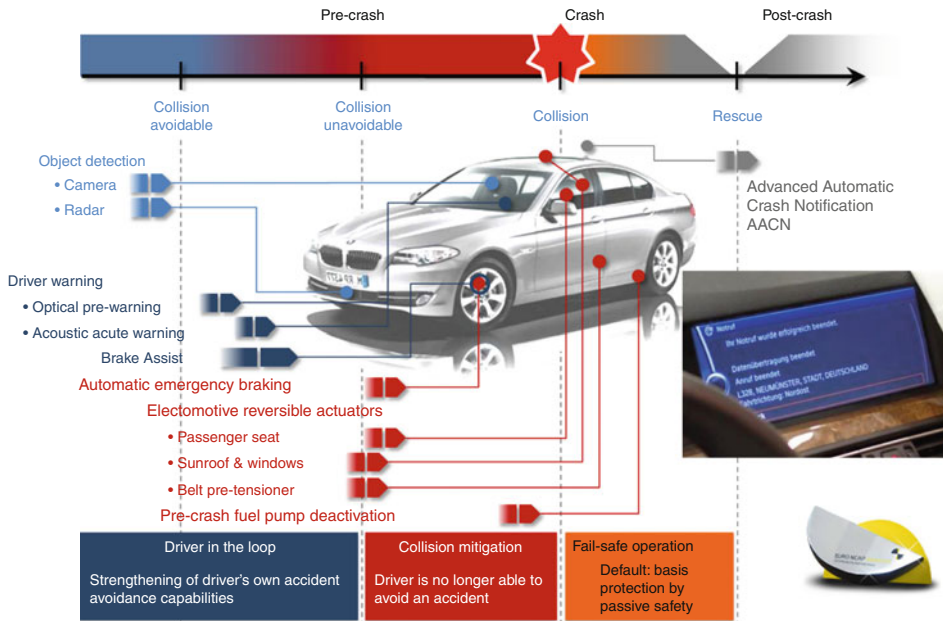
First, under certain conditions, active systems linked to the vehicle environment can allow an accident to be avoided entirely. Accident avoidance is clearly most effective for reducing deaths, injuries, and property damage. Second, linking passive and active safety can improve our ability to reduce the severity of accidents and their consequences by supporting optimization of passive safety processes.

Considerable potential for progress exists in the development and implementation of active safety systems. To an increasing extent, these systems will be capable of preventing incipient accidents during their precrash phase; hence, instead of accepting crashes (including collision trajectories and kinematics) as given and trying only to manage the consequences, the new safety paradigm will address the accident situation beginning during the precrash phase and provide strategies designed to avoid the crash entirely or reduce relative collision speeds.

Implementing this paradigm change will require meeting a number of technical challenges: Comprehensive knowledge of the vehicle's state, the driver's state, and the driving environment are required and must be provided reliably by sensors. If this data is available, the point of no return for an accident can be more accurately assessed, allowing improved adaptation of passive safety systems to the detailed accident situation. To achieve those benefits, a number of prerequisites need to be met.

- Both detection hardware and algorithms have improved noticeably in recent years. Yet the reliability of sensor systems analyzing the vehicle's environment is still not on a par with so-called inertial sensors in the vehicle, which measure accelerations, speeds, or wheel rotation. Continuous improvements are necessary – and foreseeable – to further optimize the ratio between necessary and unnecessary warnings or activations.
- Even with support by ADAS, the driver will continue to play an essential role in accident avoidance and mitigation; in some critical cases, integral safety systems will leave the final decision to the driver. Hence, increased acceptance of ADAS, better familiarity with functionalities, and improved adaptation by drivers would enhance the active safety performance of ADAS.
- Regulatory requirements and tests for such systems urgently need to be standardized and adapted to reflect field effectiveness: System development, specification, and optimization require enormous time and effort; lack of standardization or nonrepresentative tests create the risk to manufacturers that a system may not pass a specific test procedure, thus slowing down progress unnecessarily. This acute need has been clearly identified: Standardization groups, such as vFSS-Group (advanced Forward Looking Safety Systems) (vFSS 2011), consisting of partners from the automotive industry, NGOs, test institutes, insurance companies and legal authorities, are working together to develop proposals for such requirements and test scenarios.

By linking actuators and sensors of chassis control systems and driver assistance systems to passive safety systems, the concept of integral safety could mobilize unused potential for passenger protection. If a crash is unavoidable, the integral safety concept offers the opportunity to mitigate the consequences of the impending crash by targeted strategies: reduction of collision energy by braking, pretensioning restraint systems, triggering of active elements, etc. These strategies can be adapted to improve the effectiveness of airbags and seatbelts. The integral approach could also establish novel working principles for restraint systems and create new requirements on airbags and seatbelts, for example, ventilation, behavior during the ignition and expansion phases, etc.



■ Fig. 27.5

Integral safety – maximum vehicle safety by teamwork of all functions

Figure 27.5 illustrates the sequence of events relating to a crash: normal driving, the precrash phase, the crash itself, and the postcrash phase. Various systems are already available in premium vehicles that address safety throughout this entire chain of events and can protect passengers and other road users in the case of an unavoidable accident.

Support of the driver by appropriate information begins during normal driving. Camera- or radar-based sensors detect objects and other vehicles in front of the car and support the driver in maintaining a safe distance or take over monotonous tasks such as stop-and-go driving in a jam, etc.

In addition, even when these systems are not actively performing driver assistance, they can still monitor the traffic situation in front or around the vehicle in order to detect critical situations and get the driver back into the control loop. Thus, a driver whose attention has lapsed can be given optical and acoustic warnings in the case of an impending rear-end collision. In this way, the driver can react in time to avoid a collision. If he decides to execute an emergency braking maneuver, he will be effectively supported by prefilled brakes.


If there is no driver response despite the warning, or if the response occurs too late, then a collision can no longer be avoided entirely. Nonetheless, even during this phase it is still possible for active safety systems to help decrease the crash severity and the resulting risk to passengers and other road users. There are already vehicles on the road that can brake automatically in case of an unavoidable collision. Seats can be moved


into optimal position for passenger protection, and seatbelts can be pretensioned to remove slack prior to the crash.

During the crash itself, passive safety devices operate; these have already attained a high level of effectiveness. Even after the crash, vehicle-based systems lead to faster and more effective rescue and trauma care of accident victims. Sensors measure the severity, direction, and forces of the crash, count the passengers, and determine the risk of life-threatening injuries. The precise GPS coordinates of the vehicle as well as other accident data are transmitted to rescue headquarters, which can order emergency management services, arrange helicopter transport to an advanced trauma center, and initiate contact with the passengers.

As shown in the previous section, the development of passive safety is approaching a point of diminishing returns. The limitations can be seen in the case of pedestrian protection. While current assessment procedures award a favorable pedestrian safety rating based on passive safety measures alone, their benefit for pedestrians in real-life accidents is expected to be marginal and may even be difficult to detect in future accident statistics.

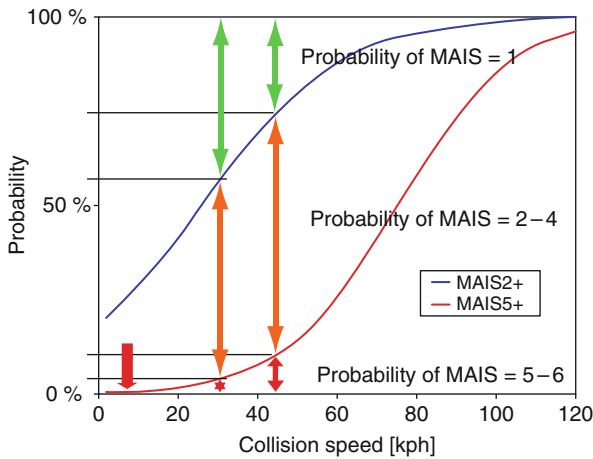
An integral safety approach appears more promising: even though it is unlikely that all collisions between automobiles and pedestrians can be avoided, the combination of avoidance of some collisions, reduction of impact velocities in many further collisions, and adequate passive safety promises far better results.

As seen previously in  Fig. 27.2, at least in certain countries pedestrians have benefited from safety measures comparably to well-protected vehicle occupants. One positive factor is the general improvement in traffic infrastructure including pedestrian friendly infrastructure, for instance in Europe. Clearly marked sidewalks, pedestrian crossings, bridges, or traffic lights provide evident benefits. A second positive factor is vehicle technology: efficient brake systems and tires enable modern vehicles to reach decelerations of more than 1 g. Brake assistants provide additional braking force so that even hesitant drivers can rapidly achieve high decelerations. A substantial number of accidents including those involving vulnerable road users have been avoided in this way.

At the same time, this discussion illustrates that an integral safety approach promises considerably greater success in injury reduction. In pedestrian collisions, the collision speed has a strong influence on injury severity. If collision speeds can be reduced in a large percentage of accidents, the protective effect as a whole will exceed that possible by passive safety measures.  Figure 27.6 displays the so-called injury risk function for pedestrian accidents, that is, the dependence of injury risk (expressed in MAIS classes) on collision speed. A decrease of the collision speed from 45 to 30 km/h considerably reduces the probability of a MAIS 5–6 injury (the most severe).

Integral safety seeks to attain a higher effectiveness than possible by passive measures alone by a holistic approach, combining different domains of development and utilizing communication among vehicles as well as connections between vehicles and their environment. However, the overall safety performance of the vehicle cannot be assessed by standard crash test procedures as carried out in the past. For example, protective preconditioning strategies affecting the vehicle or the passengers in case of a detected impending collision are ignored in such simple crash tests.

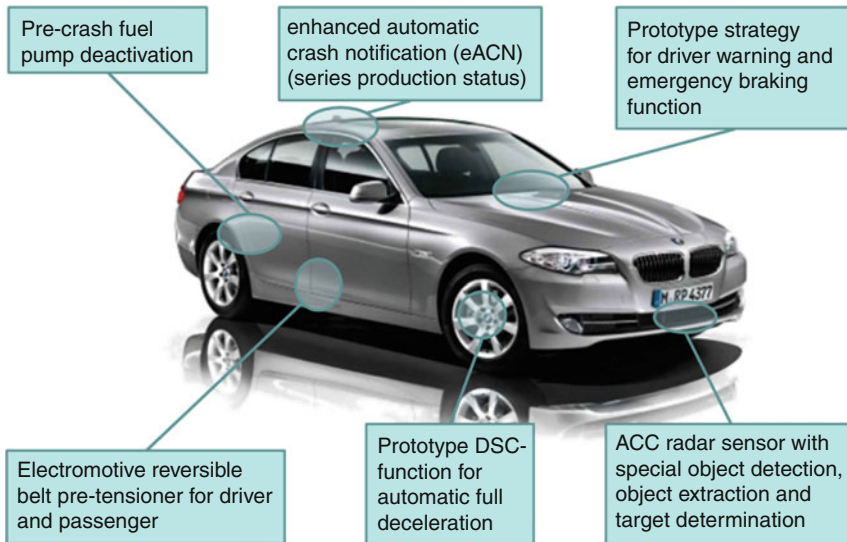
ACEA data analysis: equal effectiveness study



- Pedestrian detection by sensors.
- Increased driver awareness by warning.
- Reduction of impact speed by driver triggered emergency braking (Hydraulic Brake Assist).
- Reduction of impact speed by autonomous braking, when an impact is unavoidable.

■ Fig. 27.6

Pedestrian protection by impact mitigation by reduction of speed



■ Fig. 27.7

Prototype test vehicle – similarities and changes to series production BMW 530d

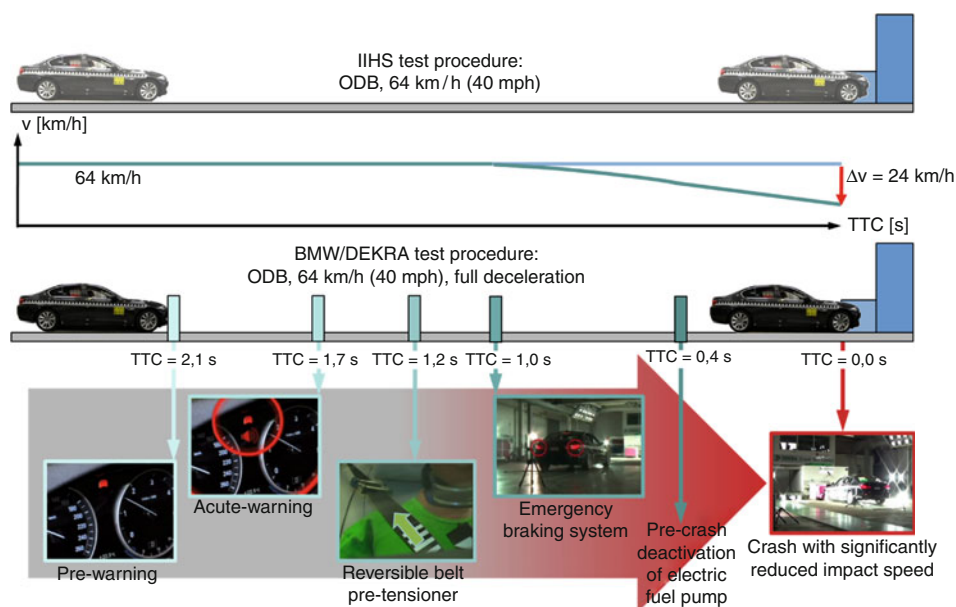
A method for including preconditioning systems in the assessment of safety performance by a crash test was demonstrated in a cooperative effort between BMW and Dekra (► Fig. 27.7). The test was performed on a specially prepared vehicle in which all systems operated autonomously. This test required substantial modifications compared to a stock

vehicle, since special measures were needed to allow target detection by a radar sensor within the test hall. In addition, the test rig needed to be equipped with a mechanism for detecting the braking vehicle and exerting only the forces required to guide the vehicle along the planned trajectory without counteracting the vehicle's own braking acceleration.

➤ *Figure 27.8* summarizes the sequence and times of vehicle actions. Following the warnings to the dummy driver – who did not respond, as a real driver might have done – the vehicle braked automatically. This emergency braking led to a collision speed of 40 km/h instead of the 64 km/h collision speed that would have otherwise occurred. This decrease resulted in a substantial reduction of occupant loads. ➤ *Figure 27.9* summarizes the test results.

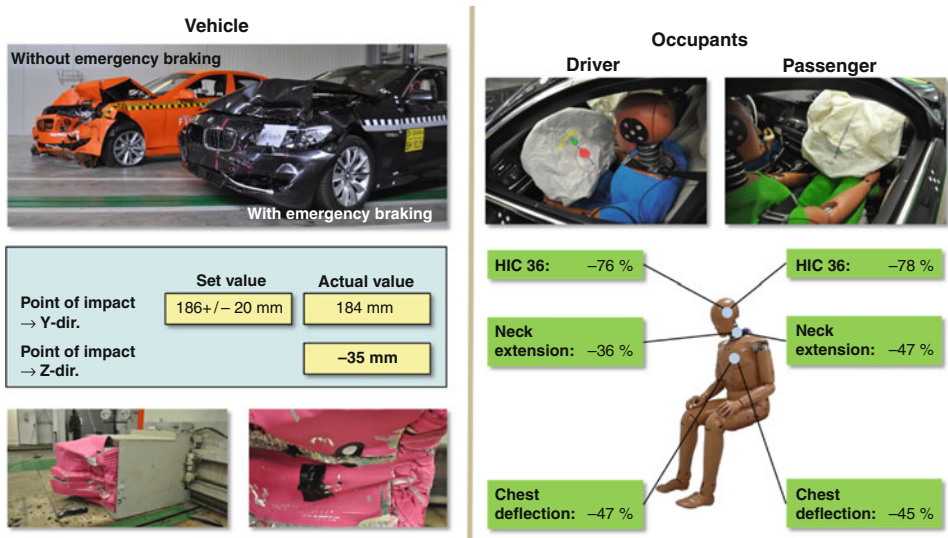
Note that the development of such systems requires enormous effort. Because accidents in real traffic are complex occurrences, at present it is only possible to address particular accident scenarios by systems designed to avoid collisions or reduce their severity. Although the limited spectrum of scenarios that can be addressed is gradually expanding due to increasingly sophisticated and reliable measurement devices and sensors, passive safety measures will continue to play an essential role for some time to come.

Integral vehicle safety comprises several stages. During the first stage, the goal is to keep the driver in the loop – or bring him back into the loop, if need be: for example, warning and evoking a reaction in case of a lapse of attention in a hazardous situation. During the next stage, the vehicle should support the responsive driver and appropriately focus or amplify his response in order to utilize the full potential of vehicle maneuvering, particularly braking.



■ Fig. 27.8

Test overview – course of events compared to standard test



■ Fig. 27.9
Test results of crash test with emergency braking reaction of the test vehicle – significant reduction of occupant loads

If (and only if) the driver fails to respond or responds far too late to avoid a collision, the vehicle can react automatically with the goal of reducing collision energy as well as preparing the vehicle and occupants for the crash in order to reduce the resulting occupant loads. Passive safety operates in a fail-safe mode to protect the occupants or other road users. Immediately following an accident, emergency rescue services should be notified by a reliable, fast, targeted, informative, automatic electronic calling system, resulting in rapid transportation to an appropriate trauma center of victims whose lives are at stake.

The development of integral safety concepts requires considerable effort. In order to make sure that all measures implemented in vehicles really achieve the highest possible safety benefits in real accidents, detailed analysis capable of predicting and optimizing these safety benefits prior to full-scale implementation are needed. Public authorities responsible for safety regulations and other agencies involved in assessing vehicle safety should define “holistic” goals and ratings: vehicle safety regulations and assessment should be oriented toward the effectiveness of a vehicle’s safety systems, considered as a whole, in reducing mortality and injuries in the field. Ideally, it should be possible to define standards for predicting effectiveness.

4 Quantification of Field Effectiveness

Benefits of improved traffic safety often accrue not only to the individual customer, but also to occupants of other vehicles, to vulnerable road users, or to the society as a whole, for example, by reduced social and health care costs, etc. Hence, in addition to automobile

manufacturers, suppliers, and customers, the public policy and opinion makers, consumer groups, as well as regulatory agencies are the key stakeholders in the introduction of safety systems. Though the interests of specific stakeholders rarely coincide entirely – for example, the benefits of a pedestrian warning system are certainly perceived more strongly by pedestrians than by drivers (one should not discount though, that a pedestrian collision is one of the worst emotional experiences that a driver can have) – there is a common interest in standardizing the assessment process for vehicle safety systems and in objectively quantifying their benefits, preferably in a way that allows individual customers and other stakeholders to evaluate and compare the potential safety benefits of vehicle-based safety systems, independent of whether passive, active, or integral safety concepts are implemented.

At present, officially sanctioned test procedures define effectiveness of passive safety systems as specified in legal regulations for consumer protection. However, international harmonization of requirements has not been achieved; a vehicle manufacturer producing for the global automobile market must absolve a large number of different tests, usually crash tests in which the vehicle is subjected to defined accident scenarios. Precisely defined measurement positions and passenger load limits provide information on the quality of passive safety in these specific test cases. Some of these test procedures, such as the Euro-NCAP Offset Crash, are rather well known to customers.


There are also defined test procedures for “classical” active safety disciplines such as stability or other areas of driving safety; in such tests, the driving stability of the vehicle or the tipping tendency in rapid swerving maneuvers are tested. The tests vary from one market to the next, with the spectrum of assessment procedures ranging from static geometric assessment (e.g., ratio of center of mass to track width) to robotically controlled maneuvers.

Active safety functions derived from driver assistance and vehicle sensors directly interact with the process of human vehicle guidance. It is a particular challenge to evaluate the effectiveness of safety systems when mental processes such as perception, reactions, and decisions or physiological safety aspects are involved. The problem is that effectiveness depends not only on technical functionality, because the response and acceptance of the driver during the precrash phase – particularly regarding the human-machine interface and the intensity of interventions in automatic accident avoidance strategies – have a decisive influence on accident prevention. Nonetheless, recognized testing and assessment scenarios will soon need to be defined for these types of active safety systems just as they have been in the past for passive systems. Compared to current safety systems, the required procedures will almost certainly involve a higher level of complexity due to the larger spectrum of parameters that can influence outcome. An additional complication is testing of passive safety systems that are coupled to active safety, such as the preconditioning system described above. These integral systems promise improved effectiveness in the field, but current official testing procedures and assessment standards do not properly measure this effectiveness. An adequate test standard should take into account the information that these systems have in a real crash; for example, precrash


data that would be provided to the restraint system should be available in order for field safety performance to be adequately demonstrated.

Hence, in order for stakeholders to quantify the effectiveness of integral safety technologies and to compare them with alternatives such as passive safety approaches, an agreed-upon objective standard of quality (or “metric”) is required. The metric should reflect the true benefits (effectiveness) of the system as it will be implemented in the field.

In the case of pedestrian protection, an appropriate metric would reward accident avoidance or injury reduction, say in ISS or MAIS classes, but it might (less obviously) also reward avoidance of false positive warnings. Secondly, a methodology for actually assessing the standard of quality based on tests, data, and analysis is needed. Like any assessment methodology, the result should be reliable (repeatable and consistent across time, independent of details such as test institute and vehicle tested), valid (representative of the defined standard of quality), and, most importantly, feasible to implement.

In the past, evaluation of field effectiveness for active safety systems has often been based on retrospective analysis of observational accident statistics. It has been possible to identify the influence of some innovations (such as the introduction of seat belts) in this way. However, retrospective evaluation of accident statistics alone is inadequate for current requirements on assessing integral safety systems. To begin with, accidents are rare events, so that a long time base is required for observing statistically significant differences in a before/after comparison. Moreover, there are fundamental obstacles to assessing the safety performance of existing safety systems based purely on observational trends: observations do not constitute a controlled study, and conclusions based on observations are thus susceptible to biases, particularly *confounding*. For example, to evaluate the effect of an advanced assistance system X, one might attempt to study accident/severity statistics observed in a sample of vehicles with X versus those without X. However, drivers of vehicles with system X are not a random sample of all drivers in such a sample. Moreover, “take rates” on different safety systems are likely to be correlated, because a vehicle with system X may be more likely to contain other systems Y, Z that also influence safety. As  Fig. 27.3 illustrates, there is also considerable temporal overlap in introduction of recent safety innovations, which further confounds the retrospective attribution of safety improvement to particular single innovations.

There has been progress in quantifying safety benefits of existing ADAS using anonymous in-vehicle data to define surrogate traffic safety indicators (Kompass et al. 2007). Here, the large volume of empirical data results in representative sampling and excellent statistical power. One limitation is of course the indirect relationship between surrogate safety indicators and field effectiveness.

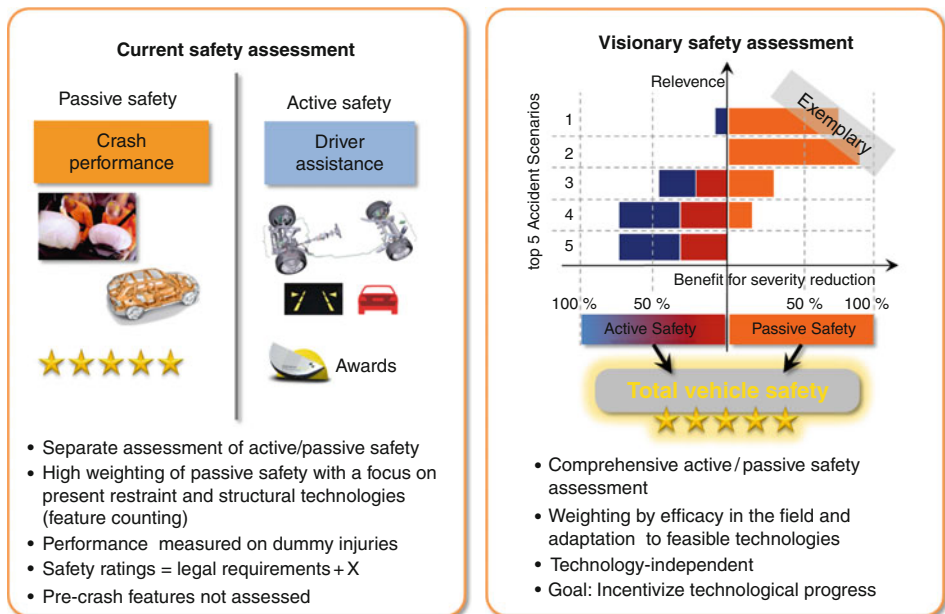
A second problem with retrospective evaluation of safety effectiveness is that the development process in novel integral safety systems usually requires optimization of so-called *operating points*, such as TTC values for which warnings are issued or interventions are carried out ( Fig. 27.8). Operating points nearly always influence field effectiveness, including both injury reduction and false positive rate. Moreover, there are usually interrelationships between operating points, sensor specifications, and field

effectiveness. If optimization were based purely on retrospective field observations, a long feedback loop between initial system configuration and adjustments due to observed field effectiveness would arise. Hence, optimization of safety systems requires algorithms capable of *predicting* safety benefits.

Considering the substantial technological effort and lead time required for implementation of possibly differing solution approaches, the above discussion indicates that reliable, valid, and feasible *safety performance prediction methodologies* accepted by all stakeholders are urgently needed (► Fig. 27.10).

A key problem for safety performance prediction in novel active or integral safety systems is that field effectiveness will increasingly involve *stochastic* variables such as the driver's perception, reactions and response to warnings; controllability of interventions; interactions with the vehicle environment (infrastructure, traffic dynamics); reactions of other road users, etc.

In principle, the stochastic element plays a role in inferring field effectiveness even from traditional crash tests, since crash tests determine dummy loads only for a limited number of discrete configurations, and the load can depend on details of crash dynamics. However, the stochastic element presents a far greater challenge in assessing the field effectiveness of active safety systems. Simply assuming a fixed value for inherently stochastic variables (e.g., assuming a reaction time of, say, 1 s) will usually lead to misleading safety predictions, because accidents are inherently rare events involving initially unlikely combinations of parameters.



■ Fig. 27.10

Vehicle safety assessment today and visionary safety assessment in the future

In trying to predict the accident probability in a driving scenario, it is very useful to imagine a large number of virtual repetitions or “realizations” of the scenario. In stochastic simulation, sometimes referred to as the “Monte-Carlo” method (a term that was coined in the 1940s by physicists, referring to the Monte-Carlo Casino in Monaco), we begin by constructing a simulation model of the scenario in a “base case” (e.g., the status quo) and then simulate many realizations of the scenario; in each individual realization, the stochastic parameters are drawn randomly from appropriate probability distributions. The resulting accident and injury probability distributions are estimated by modeling all those processes that contribute to an outcome (accident/injury or safe crossing) within the crossing scenario and repeating the simulation millions of times. Thus, if we wanted to understand and model the effect of better object visualization on reduced accident probability for vulnerable road users in a crossing scenario, we could first test and quantify how the perception probability per unit time improves (compared to the base case) due to better visualization in a representative population of drivers and vulnerable road users (with varying properties, such as lighting and contrast). In the “base case,” a driver looking ahead will perceive and correctly classify an object with a certain probability per unit time that could depend on the size of the object and its rate of change on the retina (“looming”), its position on the retina, as well as on the contrast and lighting in the environment. We could then design and simulate a virtual experiment: In the OODA paradigm (**observe, orient, decide, act**) (Green 2000), more rapid perception would increase the probability of a timely action by the driver. We could then repeat the simulation process taking improved perception into account and compare the injury statistics using the visualization system with those of the base case.

There are several useful kinds of virtual experimental designs in stochastic simulation. A “forward” simulation approach, which is currently being utilized within the development and optimization process for integral safety systems at BMW, is to begin with a scenario that is known to describe a large number of accident situations, for example, the “midblock dash” in pedestrian crossing scenarios. One then simulates an ensemble of millions of scenarios, a small percentage of which involve unfavorable outcomes such as collisions. The fidelity of the simulation in the base case is tested by comparing simulated data with empirical data. A proposed safety system is then implemented virtually, including models for its influence on all safety-relevant events and processes. As described above, this influence is reflected in a reduction in the frequency distribution of collisions and their impact characteristics.

One advantage of this “forward” simulation approach is that, particularly in many active safety concepts, warnings or interventions are triggered according to a risk analysis based, for example, on sensor inputs. In most cases of risk analysis, there is a trade-off between false positive and false negative responses that depends on the particular configuration, for example, the timing of warnings and interventions. Note that forward Monte-Carlo simulations also model those situations that would not have resulted in an accident even without the safety system, but which would have triggered the system in a particular configuration. Hence, these simulations are also capable of supporting the minimization of false positives within the optimization process.

It is also possible in principle to design a “backward” simulation approach, in which one begins from a database of real accidents and reconstructs initial situations that might have led to these accidents.

5 Summary and Conclusion

The protective potential of passive safety systems is already being exploited to a high degree. Even modest additional improvements in protection of occupants or other road users will require substantial efforts. In many cases, it is questionable whether additional passive protective measures will achieve any measurable effect at all in traffic safety. Although further developments in passive safety are possible, future challenges in vehicle safety will not be met by passive safety systems alone.

Nowadays it is possible to measure additional information from the vehicle environment by appropriate sensors and to utilize this data by networking all electronic systems in the vehicle. In particular collision scenarios, it is already possible to detect a critical situation early enough to initiate mitigating strategies or avoid an accident altogether. In case of inadequate or absent driver reactions, support is provided that allows the driver to act in an intuitively correct manner. To be sure, it is not yet feasible at present to address all possible accident scenarios by active safety systems. However, further developments will show that the driver can be provided with effective support in an increasing number of scenarios. The selection and technical design of safety systems needs to be driven by the requirements of accident scenarios in the field.

Active safety systems should focus on the driver and keep him in control of the vehicle. Support should be provided to the driver only when truly needed. Determining when support is necessary implies a certain technical robustness and reliability of active safety systems. Within the near future, only human drivers with their abilities and experience will be capable of properly interpreting and acting in complex traffic situations. Thus, certainly for the present, complex decisions should be reserved for the driver; vehicle electronics have other strengths, such as fast data processing that is not subject to fatigue. Corresponding to their particular capabilities, humans and electronic systems are ideally suited to play complementary roles in vehicle safety.

Integral safety approaches effectively combine elements of active and passive safety. In this context, evaluation of effectiveness of integral safety systems poses a particular challenge. At present, passive safety is assessed in simple test scenarios. It is tacitly assumed that these scenarios somehow represent a large fraction of critical events occurring in real traffic. Assessment of effectiveness in active, preventive, and integral protection remains difficult. The injury reduction mechanisms range from avoiding a collision entirely to reducing kinetic energy (and thus momentum transfer) or better preparing the vehicle and the passengers for a collision. Appropriate safety assessment procedures will need to take these injury reduction mechanisms properly into account. Several of the more promising approaches include simulation, particularly stochastic simulation. The challenge is to predict safety effects of new safety systems in the field reliably, in particular to

estimate directly the reduction of loads on passengers and others involved in the accident in a broad spectrum of collision scenarios. An assessment procedure for proof of effectiveness of active and preventive systems that is recognized by regulatory authorities and consumer protection organizations is urgently needed. Strong cooperation of stakeholders will be needed to develop suggestions and achieve consensus on appropriate methods and tools for reliable and valid assessment.

References

- DESTATIS (2010) Verkehrsunfälle 2009, Unfallentwicklung im Straßenverkehr. Statistisches Bundesamt, Wiesbaden
- Green M (2000) How long does it take to stop? Methodological analysis of driver perception brake times. *Transport Hum Factors* 2(3):195–216
- EURO-NCAP 2011: <http://www.euroncap.com/tests/frontimpact.aspx>
- IIHS 2011: <http://www.iihs.org/50th/default.html>
- Kompass (2011) Vehicle safety – it's time for a revolution! *Traffic Infra Tech* 1(3):58–62 December 2010 – January 2011
- Kompass K, Huber W (2009) Integrale Sicherheit – Effektive Wertsteigerung in der Fahrzeugsicherheit, VDA Technischer Kongress 2009, 25–26 March 2009, Wolfsburg
- Kompass K, Oberschaetzl M, Zollner F, Kates R (2007) Continuous, after-sales analysis of driver assistance systems. 17th ITS World Congress, Beijing
- vFSS 2011: http://www.evaluate-project.eu/pdf/evaluate-20101124-wp5-final_event_presentation_vfss.pdf
- Vision Zero Initiative 2011: <http://www.visionzeroinitiative.com>

28 Lane Change Assistance

Arne Bartels · Marc-Michael Meinecke · Simon Steinmeyer
Driver Assistance and Integrated Safety, Volkswagen Group
Research, Wolfsburg, Germany

1	Introduction	730
2	Requirements	731
3	Classification of System Functions	733
3.1	Coverage Zone Classification	733
3.2	State Diagram	735
4	Test Procedures	736
4.1	General Boundary Conditions	736
4.2	Target Vehicle Overtaking Subject Vehicle	737
4.3	Subject Vehicle Overtaking Target Vehicle	739
4.4	False Warning Test	740
4.5	Target Vehicle Moving Laterally	740
5	Sample Implementations	741
5.1	“Audi Side Assist”	742
5.2	“Lane Change Warning” from BMW	745
5.3	“Blind Spot Information System” from Ford	745
5.4	“Side Blind Zone Alert” from GM	746
5.5	“Blind Spot Monitor” from Jaguar	747
5.6	“Rear Vehicle Monitoring System” from Mazda	747
5.7	“Blind Spot Assist” from Mercedes-Benz	749
5.8	“Side Collision Prevention” from Nissan/Infiniti	750
5.9	“Blind Spot Detector” from Peugeot	751
5.10	“Blind Spot Information System” (BLIS) from Volvo	751
5.11	“Side Assist” from VW	752
5.12	Summary	753
6	Achieved Performance Capability	753
7	Further Developments	755

Abstract: More than 5% of all accidents involving injury to people take place as the result of a lane change. Therefore, it is sensible to provide the driver with a lane change assistant in order to provide support in this driving maneuver.


ISO standard 17387 “Lane Change Decision Aid System” differentiates between three different types of system: The “Blind Spot Warning” monitors the blind spot on the left and right adjacent to the driver’s own vehicle. The “Closing Vehicle Warning” monitors the adjacent lanes to the left and right behind the driver’s own vehicle in order to detect vehicles approaching from behind. The “Lane Change Warning” combines the functions of “Blind Spot Warning” and “Closing Vehicle Warning.”

Almost all major vehicle manufacturers are now offering systems that assist the driver to change lanes. Systems with “Blind Spot Warning” are available from Ford, Jaguar, Mercedes-Benz, Nissan/Infiniti, Peugeot, and Volvo. Systems with “Lane Change Warning” are available from Audi, BMW, Mazda, and VW. All vehicle manufacturers use an optical display in or near to the exterior mirrors in order to show information for the driver. The majority of vehicle manufacturers use radar sensors that are installed in the rear of the vehicle. Two-level, escalating driver information is only offered in some of the systems. The type of escalation (optical, acoustic, tactile, lateral guidance intervention) usually differs from one vehicle manufacturer to another.

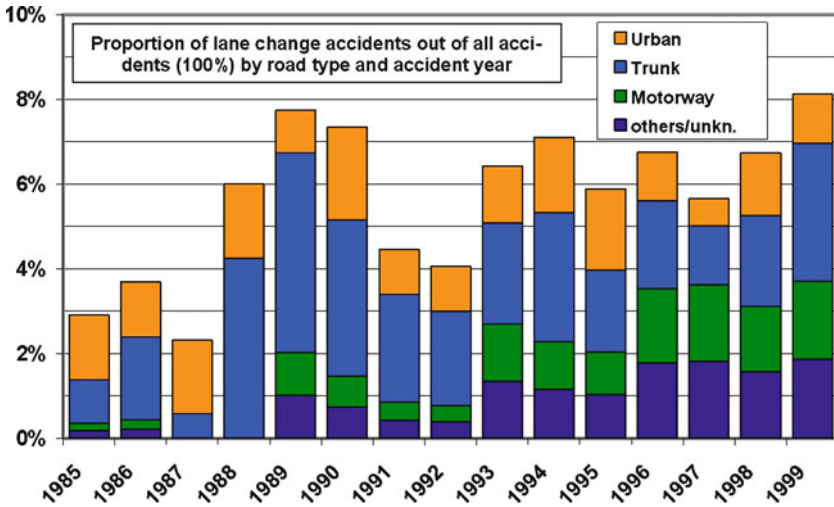
The performance capability of the lane change assistants described above is already quite considerable. However, all of these systems have their limits, and the vehicle manufacturers need to inform drivers of these in the owner’s manual, for example.

1 Introduction

The purpose of driver assistance systems is to offer the drivers additional convenience and safety by supporting them in their driving task. The customer benefit expected from a driver assistance system is particularly high if the driving task in which the driver is to be assisted is one which harbors a high potential for error. The lane change is one of these driving tasks with a high error potential.

This is indicated by a statistical analysis of accidents involving injuries to people, which have been collected in a database held at Volkswagen Accident Research and GIDAS (German In-Depth Accident Study).  [Figure 28.1](#) shows the proportion of lane change accidents with passenger cars as the main causal factor for the road types of urban, rural, and motorway for the years 1985–1999. It is clear that on average more than 5% of all accidents take place during a lane change. It is also clear that a majority of these accidents occur on trunk roads or motorways.

These accident statistics indicate the importance of providing drivers with a system that can support them when changing lanes. Initially, this support is to be configured for trunk road and motorway scenarios.



■ Fig. 28.1

Proportion of lane change accidents with passenger car as the main causal factor out of all accidents by road type and accident year (Accident database of Volkswagen Accident Research and GIDAS)

2 Requirements

The driver needs to be able to preclude any risk to other road users when making a lane change. According to the relevant regulations, it is the driver's responsibility to check the area at the rear and to the side of the vehicle before a lane change. As part of this, it is mandatory both to look in the exterior and rear-view mirrors and to glance over the shoulder. Omitting the glance over the shoulder, or if the exterior mirrors are not set correctly, or if the driver is simply inattentive, leads to the risk of failing to notice other road users in the blind spot. If a lane change maneuver is initiated under such circumstances, then this can result in a collision with the vehicle in the adjacent lane.

Another frequent cause of accidents when changing lanes is failure to estimate the speed of overtaking vehicles correctly. On motorways in particular, drivers frequently underestimate the approach speed of far distant vehicles which are, however, coming up quickly from behind. In this situation, a lane change can lead not only to a collision with the overtaking vehicle if it is unable to brake sufficiently, but also to rear-end collisions with other road users who are unable to respond in good time to the severe deceleration of the overtaking vehicle.

The driver also needs support when changing lanes toward the front passenger side. In Germany, this is mandatory due to the obligation to drive on the right. Following an overtaking maneuver, the driver is obliged to return to the right-hand lane as soon as the traffic situation permits. In contrast to this, overtaking on the front passenger side is also practiced in many other European countries. Furthermore, in the USA, it is an everyday

reality that other road users will be driving at almost the same speed in both adjacent lanes, in the blind spot of the driver's own vehicle.

This analysis leads to the following functional requirements on a lane change assistant:

- The lane change assistant should support the driver during lane change maneuvers via detecting potential hazardous situations that result due to the driver failing to monitor the area around the vehicle adequately.
- For this purpose, the assistance function should be capable of detecting road users approaching rapidly from behind as well as detecting other road users in the blind spot of the driver's own vehicle.
- The assistance function should operate for both adjacent lanes equally, on the driver's and the front passenger's sides.
- Ideally, the assistance function will be available under all road, weather, and traffic conditions with approximately the same level of quality.

Of particular importance is the human-machine interface (HMI) between the driver and lane change assistant. If the system's monitoring of the area around the vehicle indicates that the lane change is potentially dangerous, then the driver is informed about this in a suitable and timely manner. The information can in principle be delivered via the optical, acoustic, or tactile sensory channels of human beings. In configuring the HMI, however, attention should be paid to encouraging the driver to look in the mirror, as this remains an obligation on the driver even when the lane change assistant is activated. Positioning optical displays in or near to the exterior mirrors represents a solution to this requirement. The spatial proximity between the exterior mirror and optical display ensures that the driver can simultaneously perceive the optical information from the assistance function when looking in the mirror. The brightness of these optical displays should be configured so that they can be easily perceived by the driver under all ambient conditions that arise. On the other hand, neither the driver nor drivers of other vehicles are allowed to be distracted or dazzled by the optical displays, in particular at night.

When configuring the HMI, it will also be necessary to decide whether the driver information should be provided on a one- or two-level basis. In a two-level driver information system, escalation from information level 1 to information level 2 takes place as soon as the driver's intention to change lanes is detected. This escalation does not take place in a one-level driver information function.

Information level 1 shows the driver each vehicle that is a potential hazard given a lane change. This happens even if the driver is not intending to change lanes. Although the display information level 1 should be noticeable to the driver, it should cause neither interference nor distraction even when activated frequently. If the optical displays are positioned in or close to the exterior mirrors, this can be achieved, for example, by appropriate control of the lamp brightness depending on the ambient light level.

In information level 2, the driver's intention to change lanes is also detected, e.g., by actuation of the turn signal lever. If the driver intends to make a lane change, and this lane change is evaluated as potentially dangerous based on the system's perception of the surroundings, then more intensive information should be provided to the driver. In terms

of providing information for the driver by optical displays in or close to the exterior mirrors, this can be done by means of a very bright, brief flash of the optical displays, for example. Tactile or acoustic information or an intervention in the lateral guidance of the vehicle can also be used for this, the latter one, e.g., with the help of an ESP system or an electronically controllable steering actuator system.

Another important factor for the lane change assistant is an intelligent information strategy. In order to ensure adequate customer acceptance, the lane change assistant needs, on the one hand, reliably to display all traffic situations that are perceived as potentially hazardous. On the other hand, the driver must not be given unnecessary information. Unnecessary in this context includes information about a vehicle in the adjacent lane that, although detected by the environment sensors, is moving sufficiently slowly and is still far enough away to allow a lane change to be performed without risk. Another unnecessary item of information would be for a vehicle driving straight ahead in the next lane but one. The information strategy therefore needs to evaluate the measurement data from the environment sensors and, on this basis, decide very carefully whether the driver should be informed or not.

3 Classification of System Functions

The current ISO standard 17387 “Lane Change Decision Aid System” specifies various configurations of the lane change assistant and classifies them into various sub-types. Furthermore, a system status diagram is defined with system statuses and transition conditions. These are presented below.

3.1 Coverage Zone Classification

Three different system types are defined in ISO standard 17387. These differ in terms of the zones covered by the environment sensors. ● [Table 28.1](#) shows an overview.

■ Table 28.1

Classification by zone coverage (ISO 2008)

Type	Left adjacent zone coverage	Right adjacent zone coverage	Left rear zone coverage	Right rear zone coverage	Function
I	X	X			Blind Spot Warning
II			X	X	Closing Vehicle Warning
III	X	X	X	X	Lane Change Warning

The specified system types have the following functions:

- Type I systems provide a warning about vehicles in the blind spot on the left and right sides. They do not provide a warning about vehicles that are approaching from the rear on the left or right side.
- Type II systems provide a warning about vehicles that are approaching from the rear on the left and right sides. They do not provide a warning about vehicles in the blind spot on the left or right side.
- Type III systems provide a warning both about vehicles in the blind spot and vehicles approaching from the rear, in both cases on the left and right sides.

Type II and type III systems are themselves subdivided into three sub-categories. In the aforementioned standard, these are distinguished according to the maximum permitted relative closing speed of the target vehicle approaching from the rear v_{\max} , as well as the minimum roadway radius of curvature R_{\min} . [Table 28.2](#) shows an overview.

The maximum target vehicle closing speed has a direct influence on the required sensor range, given the computation time of the system and a specified minimum response time of the driver. At $v_{\max} = 20 \text{ m/s}$, a computation time of the system of 300 ms and a required minimum response time of 1.2 s, the minimum sensor range is $20 \text{ m/s} \cdot (1.2 \text{ s} + 0.3 \text{ s}) = 30 \text{ m}$. The sensor range must be increased if information is to be provided in good time at even faster approach speeds. At $v_{\max} = 30 \text{ m/s}$, this means the minimum sensor range is already 45 m.

There are two reasons for classification with regard to the minimum roadway radius of curvature. On the one hand, early detection of the target vehicle can be made more difficult by the restricted detection range of the environment sensors used. For example, given a lobe-shaped detection range, the opening aperture of the sensor is a significant factor in achieving good coverage of the relevant lane when driving on a curve. On the other hand, there is a relationship between the maximum target vehicle closing speed and the roadway radius of curvature. For a given curve radius and a typical subject vehicle speed, the closing speed of a target vehicle is limited by driving dynamics parameters.

Table 28.2
Classification by target vehicle closing speed and roadway radius of curvature (ISO 2008)

Type	Maximum target vehicle closing speed (m/s)	Minimum roadway radius of curvature (m)
A	10	125
B	15	250
C	20	500

3.2 State Diagram

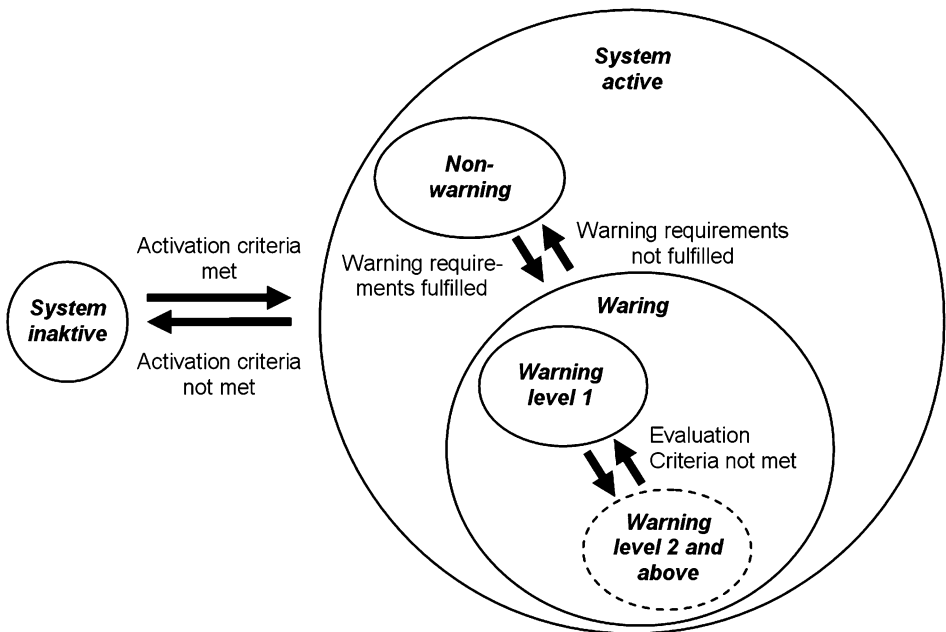
ISO standard 17387 specifies a state diagram with system statuses and transition conditions for the lane change assistant. This is shown in [Fig. 28.2](#).

In the inactive state the system shall give no information to the driver. This state may be a power off state or a ready state. In a ready state the system may detect target vehicles, but shall not issue information to the driver.

For a transition to the active state several activation criteria may be used at the same time. For example, the system may be activated if the ego vehicle is driving faster than a specified minimum activation speed and if the driver is pressing a button simultaneously. The system may be deactivated, if the driver presses the off button or is driving below the minimum activation speed.

In the active state information is only given to the driver if certain requirements are fulfilled; for example, a vehicle is detected in the blind spot or else a vehicle is approaching from the rear at high speed. No information is given to the driver if these requirements are not fulfilled.

The information can be given to the driver using several distinct levels. Information level 1 involves “discreet” information to the driver at a lower level of urgency than level 2 driver information. The driver is given level 2 information if certain evaluation criteria



■ Fig. 28.2

State diagram for a lane change assistant acc. to ISO 17387 (ISO 2008)

are met, indicating that it is the driver's intention to make a lane change. These selection criteria can include:

- (a) Activation of the turn signal lever
- (b) Evaluation of the steering angle or steering torque
- (c) Evaluation of the ego vehicle position or ego vehicle motion within the lane
- (d) Evaluation of the lateral distance from a vehicle in the adjacent lane

In case of (c), it is possible to use synergy effects with a system that is possibly present for detecting the lane markings.

ISO standard 17387 also describes test cases for activation of driver information in detailed and quantitative terms. Some of these are presented in the following section.

4 Test Procedures

Test procedures that are standard and applicable to all assistance systems to be assessed must be defined in order to judge the functional capability of a lane change assistant objectively. The boundary conditions under which these tests are performed must be specified and complied with as precisely as possible.

This is done in ISO standard 17387 for all three functions dealt with in the standard:

- Blind Spot Warning
- Closing Vehicle Warning
- Lane Change Warning

This test procedure is described below taking the example of the “Blind Spot Warning” function. After general boundary conditions have been defined, the following test cases are considered for this purpose:

- Target vehicle overtaking subject vehicle
- Subject vehicle overtaking target vehicle
- False warning test
- Subject vehicle moving laterally

The systematic approach taken in the ISO standard should be made clear on the basis of these test cases. Complete classification of a lane change assistant is not possible with this brief excerpt from the standard. The complete ISO standard 17387 would have to be used for this purpose.

4.1 General Boundary Conditions

❖ **Table 28.3** below shows all relative boundary conditions for performing the test. The subject vehicle (line 1) in this case is the vehicle equipped with the lane change assistant. The target vehicle (line 2) in this case is the vehicle which is approaching the subject vehicle from behind and is located in its blind spot. The properties with regard to the road,

■ Table 28.3


General boundary conditions for performing the tests

1	Subject vehicle	<ul style="list-style-type: none"> • Driving straight ahead • Speed: ≥ 20 m/s 			
2	Target vehicle	<ul style="list-style-type: none"> • Motorcycle with rider • Length: 2.0 m ... 2.5 m • Width: 0.7 m ... 0.9 m (not including side mirrors) • Height: 1.1 m ... 1.5 m (not including windscreen) 			
3	Ambient conditions	<ul style="list-style-type: none"> • Road: flat, dry asphalt, or concrete surface • Ambient temperature during the test: $10^{\circ}\text{C} \pm 30^{\circ}\text{C}$ • Horizontal visibility: > 1 km 			
4	Measuring system	<ul style="list-style-type: none"> • Distance < 2 m • 2 m ... 10 m • > 10 m 	<ul style="list-style-type: none"> • Measuring accuracy ≤ 0.1 m • $\leq 5\%$ • ≤ 0.5 m 	<ul style="list-style-type: none"> • Time < 200 ms • 200 ms ... 1 s • > 1 s 	<ul style="list-style-type: none"> • Measuring accuracy ≤ 20 ms • $\leq 10\%$ • ≤ 100 ms


temperature, and visibility conditions are specified in line 3. The reaction times of the driver assistance system as well as the longitudinal and lateral distances between the subject vehicle and the target vehicle must be measured using a separate measuring system. This must be completely independent from the driver assistance system. The requirements in terms of the measuring accuracies of this measuring system are shown in line 4.


4.2 Target Vehicle Overtaking Subject Vehicle

The purpose of this test is to check that the blind spot warning system gives warnings when required as the target vehicle overtakes the subject vehicle.

The basic sequence of the test is shown in  Fig. 28.3. The test target vehicle (1) approaches the subject vehicle (4) at a differential speed from 1 m/s to 3 m/s. The lateral distance between both vehicles (2) in this case is 2.0 m to 3.0 m, measured between the outermost edge of the subject vehicle's body (excluding the exterior mirror) and the centreline of the test target vehicle.

Line A and line B are located at a distance of 30.0 m and 3.0 m behind the trailing edge of the subject vehicle, respectively. Line C is approximately at the shoulder level of the driver (specifically: center of the 95th percentile eyellipse). Line D is located at the leading edge of the subject vehicle.

As the test target vehicle approaches and overtakes the subject vehicle, the system shall meet the following test requirements specified in  Table 28.4.

This test shall be repeated according to  Table 28.5 for a total of 12 trials. During night time conditions, no illumination shall be provided other than the standard headlamps and tail lamps of the subject vehicle and test target vehicle. If it can be shown that the ambient light conditions have no effect on the system's performance, then the tester may choose to perform either the daytime tests or the night time tests for a total of six trials.

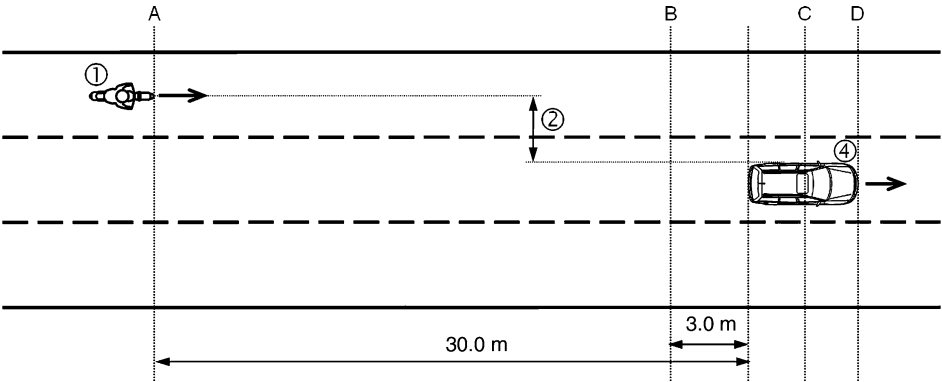


Fig. 28.3
Test scenario “target vehicle overtakes subject vehicle” (ISO 2008)

Table 28.4
Test requirements for the “target vehicle overtaking subject vehicle” scenario

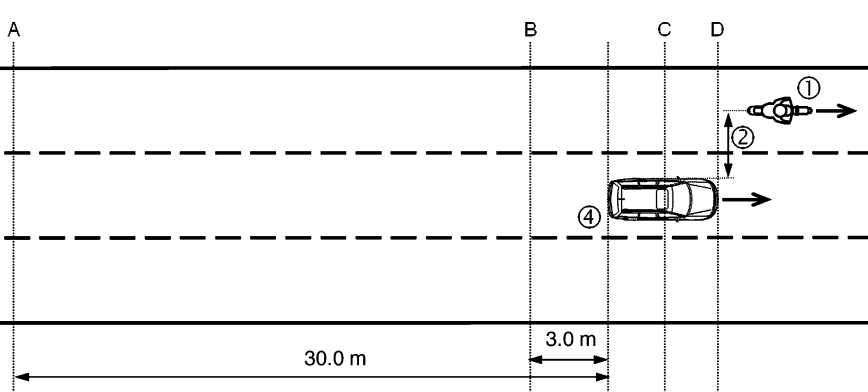
Target vehicle	System response
Behind line A	<ul style="list-style-type: none">• Shall give no warnings
Crosses line A	<ul style="list-style-type: none">• Shall initiate a warning• On the correct side of the subject vehicle
Crosses line B	<ul style="list-style-type: none">• Shall initiate a warning• On the correct side of the subject vehicle• No later than the time at which the leading edge of the test target vehicle crosses line B plus the system response time
Crosses line C	<ul style="list-style-type: none">• Shall sustain the warning• At least until the leading edge of the test target vehicle crosses line C
Forward of line D	<ul style="list-style-type: none">• Shall terminate the warning• No later than after the time at which the trailing edge of the test target vehicle crosses line D plus the system response time

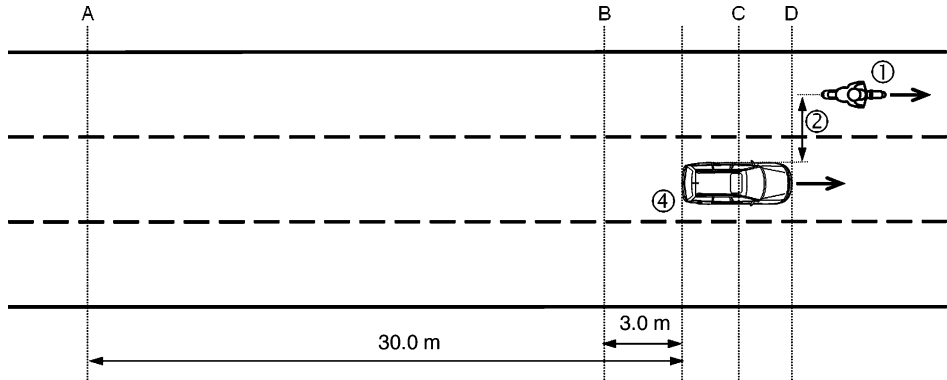
Table 28.5
Test trials for the “target vehicle overtaking subject vehicle” scenario (ISO 2008)

	Day (trials)	Night (trials)
Target vehicle to the left of subject vehicle	3	3
Target vehicle to the right of subject vehicle	3	3

4.3 Subject Vehicle Overtaking Target Vehicle

The purpose of this test is to check that the blind spot warning system gives warnings when required as the subject vehicle overtakes the target vehicle.

The basic sequence of the test is shown in  Fig. 28.4. The subject vehicle (4) overtakes the target vehicle (1) at a differential speed from 1 m/s to 2 m/s. The lateral distance



■ Fig. 28.4

Test scenario “subject vehicle overtaking target vehicle” (ISO 2008)

■ Table 28.6

Test requirements for the “subject vehicle overtaking target vehicle” scenario

Target vehicle	System response
Forward of line D	<ul style="list-style-type: none"> ● Shall give no warnings
Crosses line D	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the correct side of the subject vehicle ● After the trailing edge of the target vehicle crosses line D
Crosses line C	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the correct side of the subject vehicle ● No later than after the time at which the leading edge of the test target vehicle crosses line C plus the system response time, plus the optional warning suppression time
Crosses line B	<ul style="list-style-type: none"> ● Shall sustain the warning ● At least until the leading edge of the test target vehicle crosses line B
Crosses line A	<ul style="list-style-type: none"> ● Shall terminate the warning ● No later than after the time at which the leading edge of the test target vehicle crosses line A plus the system response time

between both vehicles (2) in this case is once again 2.0–3.0 m, measured between the outermost edge of the subject vehicle's body (excluding the exterior mirror) and the centerline of the test target vehicle. The distances between the lines are as defined in the previous test scenario.

As the subject vehicle approaches and overtakes the target vehicle, the system shall meet the following test requirements specified in [Table 28.6](#).

This test should be performed 12 times in total acc. to [Table 28.5](#), including the requirements defined there with regard to lighting devices and ambient light.

4.4 False Warning Test

The purpose of this test is to check that the blind spot warning does not give warnings when the test target vehicle is in a lane beyond the adjacent lane. The sequences of tests described above shall be repeated with the following modifications:

- In each test trial, the lateral distance between the outermost edge of the subject vehicle's body (excluding the exterior mirror) and the centerline of the test target vehicle shall be maintained at 6.5–7.5 m.
- The system shall give no warnings during these test trials.

4.5 Target Vehicle Moving Laterally

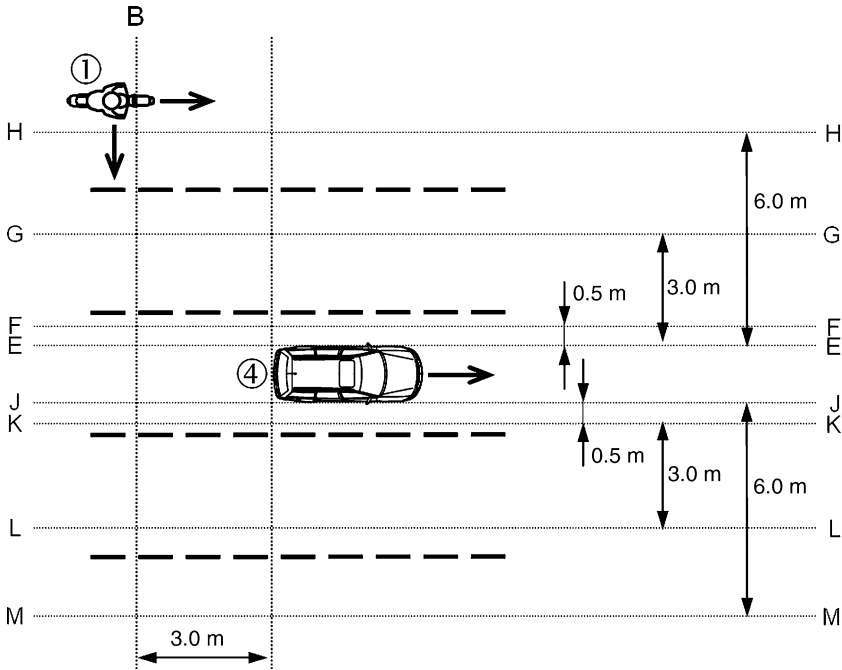
The purpose of this test is to check that the blind spot warning system gives warnings when required as the target vehicle moves laterally near the subject vehicle.

The basic sequence of the test is shown in [Fig. 28.5](#). The test target vehicle shall match the subject vehicle's speed such that the leading edge of the test target vehicle (1) is positioned between line B and the subject vehicle (4) throughout the test. To begin the test, the test target vehicle shall be completely to the left of lines H. The test target vehicle shall move toward the right at a lateral speed of 0.25–0.75 m/s until it is completely to the right of line M. Then the target vehicle shall move toward the left, once again at a lateral speed of 0.25–0.75 m/s, until it is once again completely to the left of line H.

As the test target vehicle moves from left to right, the system shall meet the following test requirements specified in [Table 28.7](#).

As the target vehicle moves from right to left, the system shall meet the following test requirements specified in [Table 28.8](#).

This test should be performed six times in total acc. to [Table 28.9](#). The same requirements apply to lighting devices and ambient light as for [Table 28.5](#).



■ Fig. 28.5

Test scenario “target vehicle moving laterally” (ISO 2008)

5 Sample Implementations

Driver assistance systems which support the driver during lane change maneuvers have been available from several vehicle manufacturers for several years now. Initially, these were used in upper category vehicles: at Audi in the A8 and Q7, at BMW in the 7-Series, at Mercedes in the S Class, and at VW in the Phaeton and Touareg. In the meantime, however, a democratization of this driver assistance system has become apparent. To an increasing extent, lane change assistance systems are also to be found in medium category vehicles such as the Audi A4 or the Volvo S40. Indeed, the lane change assistant is now also available in lower vehicle categories such as the Mazda 3.

The characteristics of the systems from individual vehicle manufacturers differ markedly from one another in some cases, although the majority of them can be assigned to the various categories of ISO 17,387, as was described in [Sect. 3](#). The differences chiefly concern the various sensors that are used for environment perception, as well as the method by which the driver is informed.

The different vehicle manufacturers have each chosen different product names in order to differentiate themselves from their competitors, and also certainly to emphasize

■ Table 28.7

Test requirements for the “target vehicle moving laterally” scenario

Target vehicle	System response
Left of line H	<ul style="list-style-type: none"> ● Shall give no warnings
Crosses line H	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the left side of the subject vehicle
Crosses line G	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the left side of the subject vehicle ● No later than the time at which the right edge of the test target vehicle crosses line G plus the system response time
Crosses line F	<ul style="list-style-type: none"> ● Shall sustain the warning ● At least until the right edge of the test target vehicle crosses line F
Right of line E	<ul style="list-style-type: none"> ● Shall terminate the warning ● No later than the time at which the left edge of the test target vehicle crosses line E plus the system response time
Between lines E and J	<ul style="list-style-type: none"> ● Shall give no warnings
Crosses line J	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the right side of the subject vehicle
right of line K	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the right side of the subject vehicle ● No later than the time at which the left edge of the test target vehicle crosses line K plus the system response time
Right of line L	<ul style="list-style-type: none"> ● Shall sustain the warning ● At least until the left edge of the test target vehicle crosses line L
Right of line M	<ul style="list-style-type: none"> ● Shall terminate the warning ● No later than the time at which the left edge of the test target vehicle crosses line M plus the system response time

the proprietary system functions. For example, Audi calls its lane change assistance “Audi Side Assist”; almost the same system in use at VW is called “Side Assist.” Its name at BMW is “Lane Change Warning.” Mercedes Benz calls its system “Blind Spot Assist.” Mazda, for its part, uses the name “Rear Vehicle Monitoring System.” Volvo has called its system the “Blind Spot Information System.”

The systems from the individual vehicle manufacturers will be presented briefly below in alphabetical order. This will focus on differences between functions and the HMI.

5.1 “Audi Side Assist”

“Audi Side Assist” informs the driver both about vehicles in the blind spot and vehicles that are approaching quickly from behind. This information is given both on the driver’s and front passenger’s sides.

■ Table 28.8

Test requirements for the “target vehicle moving laterally” scenario

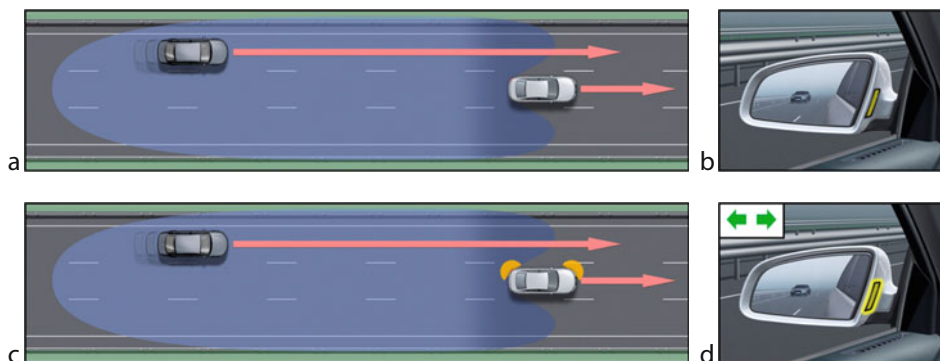
Target vehicle	System response
Right of line M	<ul style="list-style-type: none"> ● Shall give no warnings
Crosses line M	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the right side of the subject vehicle
Crosses line L	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the right side of the subject vehicle ● No later than the time at which the left edge of the test target vehicle crosses line L plus the system response time
Crosses line K	<ul style="list-style-type: none"> ● Shall sustain the warning ● At least until the left edge of the test target vehicle crosses line K
Left of line J	<ul style="list-style-type: none"> ● Shall terminate the warning ● No later than the time at which the right edge of the test target vehicle crosses line J plus the system response time
Between lines E and J	<ul style="list-style-type: none"> ● Shall give no warnings
Crosses line E	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the left side of the subject vehicle
Left of line F	<ul style="list-style-type: none"> ● Shall initiate a warning ● On the left side of the subject vehicle ● No later than the time at which the right edge of the test target vehicle crosses line F plus the system response time
Left of line G	<ul style="list-style-type: none"> ● Shall sustain the warning ● At least until the right edge of the test target vehicle crosses line G
Left of line H	<ul style="list-style-type: none"> ● Shall terminate the warning ● No later than the time at which the right edge of the test target vehicle crosses line H plus the system response time

■ Table 28.9

Test trials for the “target vehicle moving laterally” scenario (ISO 2008)

Day	Night
3 Trials	3 Trials

“Audi Side Assist” is based on two 24 GHz radar sensors with narrow-band transmission that are integrated behind the left and right corners of the rear bumper, making them invisible from the outside. The range of the sensors was restricted to 50 m in the first generation, whereas the current version achieves ranges from 70 m to 100 m. This means the driver is informed in good time about vehicles that are approaching at high speed from behind. The side areas to the left and right of the ego vehicle are scanned by a pronounced and specifically expanded side lobe of the radar sensors. This means information can also



■ Fig. 28.6

“Audi Side Assist” (Manual 2008). (a), (c) Detection range. (b), (d) Integration of the light in the exterior mirror housing. (a), (b) Continuous activation of the yellow indicator if lane change is critical. (c), (d) Brief, bright flashing of the yellow indicator if the turn signal is activated and the lane change is critical

be provided about vehicles in the blind spot. More details about “Audi Side Assist” can be found in (Popken 2006).

The driver information is given by lights integrated in the housing of the left and right exterior mirrors. The lamp in question lights up if the lane change appears potentially hazardous based on the situation perceived by the system (see ► Fig. 28.6a, b). This information level 1 is configured to be subliminal, i.e., the driver only notices it by looking directly into the mirror. This means the function of the system can be continuously experienced by the driver even in situations that are not potentially dangerous, without however the driver being subject to interference or distraction due to the light. Information level 2 is activated if the driver operates the turn signal. In this case, the driver is informed about a hazard involved in a lane change by the light flashing brightly several times (● Fig. 28.6c, d). If the turn signal is activated continuously then detected vehicles are indicated by the lamp remaining continuously lit; there is no continuous flashing. Additional documents about the HMI of “Audi Side Assist” can be found in (Vukotich et al. 2008).

The radar sensors of “Audi Side Assist” are referred to as narrow-band systems within the specifications of the ISM band between 24.000 GHz and 24.250 GHz. Their transmission power of maximum 20 dBm EIRP complies with European standard EN 300 440. These radar sensors do not require special modification of the radio certification regulations. They are not subject to the restrictions imposed on 24 GHz radars with wide-band transmission, and do not have to be switched off in the vicinity of radio-astronomical facilities. Sensors similar to those used by Audi are also employed by BMW, Mazda, and VW.

The first “Audi Side Assist” systems had an activation speed of 60 km/h. Current systems can be used in full at a lower speed, above 30 km/h. This means “Audi Side Assist” can be used both on motorways and trunk roads, as well as in urban areas.

“Audi Side Assist” was offered for the first time in 2005 in the Audi Q7. Nowadays, the system is available in many Audi vehicles. The additional price for the system, which is offered as an optional extra, is currently 550 € in the Audi A4, S4, A5, S5, Q5, A6, and S6 or 600 € in the Audi Q7, A8, and S8, in each case including VAT.

5.2 “Lane Change Warning” from BMW

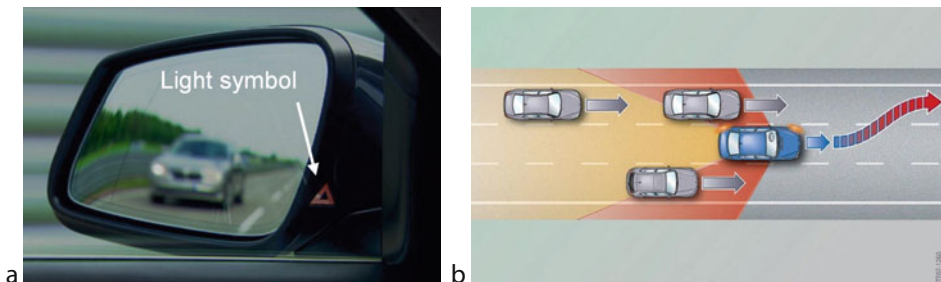
The “Lane Change Warning” from BMW assists the driver by permanently observing the adjacent lanes from a speed of 50 km/h and above; it gives a warning of a potentially dangerous situation for overtaking and lane change maneuvers. If there is a vehicle in a critical area, this is indicated by a permanently illuminated symbol in the housing of the exterior mirror (see ► Fig. 28.7a). If the driver sets the turn signal for a lane change in this situation, the illuminated symbol begins to flash and the steering wheel starts to vibrate. This ensures additional attention, and increases the warning effect (Ehmanns et al. 2008).

The Lane Change Warning has two 24 GHz radar sensors on both sides to detect other road users to the rear and sides of the driver’s own vehicle, as well as objects that are in the blind spot (see ► Fig. 28.7b). The sensors have a range up to 60 m. Their function is largely unaffected by weather conditions, and they are integrated invisibly in the rear of the vehicle. The range of the sensors means that the driver can also be warned in good time even given relatively fast approach speeds. Similar sensors to those used by BMW are also employed by Audi, Mazda, and VW.

BMW currently offers the “Lane Change Warning” in the 7-Series and the 5-Series saloon for 650 €; customers can purchase the system in the 5-Series Grand Turismo for 620 €, in each case including VAT.

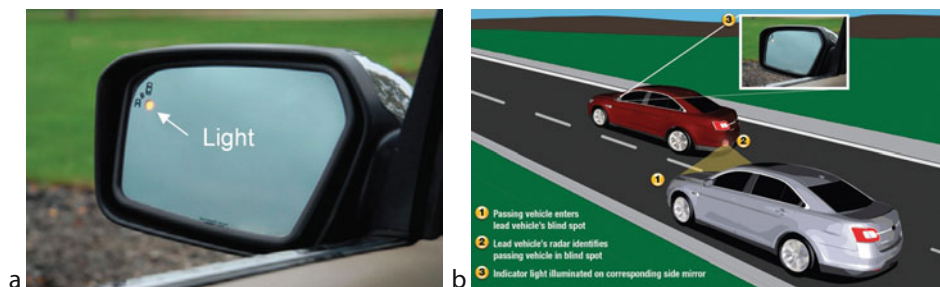
5.3 “Blind Spot Information System” from Ford

Ford’s “Blind Spot Information System” (BLIS®) is a feature that helps detect vehicles in blind spots during normal driving. The feature uses two multiple-beam radar modules,



■ Fig. 28.7

“Lane Change Warning” from BMW (Forum 2008). (a) Red light symbol in the exterior mirror housing. (b) Monitored area



■ Fig. 28.8

“Blind Spot Information System” from Ford (Ford 2009). (a) Yellow light behind the mirror glass of the exterior mirror. (b) Schematic diagram of function

which are packaged in the rear quarter panels – one per side. The radar identifies when a vehicle (truck, car, motorcycle, bicycle, etc.) enters the defined blind spot zone (see ● Fig. 28.8b) and illuminates a yellow indicator light on the corresponding sideview mirror (see ● Fig. 28.8a), providing a warning that a vehicle is approaching. According to Ford, the “Blind Spot Information System” provides additional visibility and therefore reduces stress on the driver, as well as increasing safety in road traffic.

The area monitored on the driver’s and front passenger’s sides is approximately 3 m wide and extends from the rear-view mirror backward to approximately 3 m behind the rear of the vehicle (see ● Fig. 28.8b). This means vehicles approaching at speed from behind cannot be detected within sufficient time. As a result, the customer benefit is restricted, especially on motorways.

The warning lights are integrated behind the exterior mirror glass (see ● Fig. 28.8a). The radar sensors are positioned at the sides of the rear bumper (see ● Fig. 28.8b). They operate in the frequency range around 24 GHz. The side area is monitored using several pronounced antenna lobes, allowing the bearing of the objects to be assigned. The activation speed of the system is 10 km/h.

In Europe, Ford is currently offering the “Blind Spot Information System” in the Galaxy and S-MAX models for an additional price of 495 € including VAT. In North America, it is offered in the Taurus Limited or the Ford Fusion as part of an equipment package including other features, for \$2,000 or \$1,795, respectively.

5.4 “Side Blind Zone Alert” from GM

If the “Side Blind Zone Alert” system from GM detects a vehicle in the adjacent lane, the system will illuminate a small orange icon on the side view mirror alerting the driver to a potential collision (see ● Fig. 28.9a). If the driver activates the turn signal in the direction of the detected vehicle, the symbol will flash. Otherwise it remains illuminated until the other vehicle has left the blind zone. According to GM, this technology makes lane changes safer because it alerts driver’s to vehicles that otherwise might escape their vision.

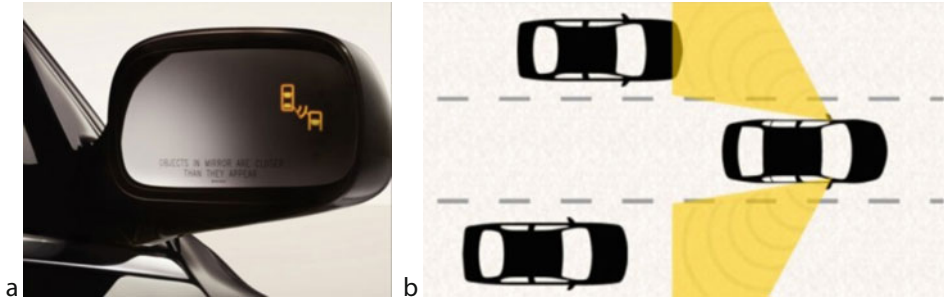


Fig. 28.9

“Side Blind Zone Alert” from GM (GM 2010). (a) Orange light icon behind the mirror glass of the exterior mirror. (b) Schematic diagram of the function with monitoring area

The system uses alternating radar beams that sweep outward from the vehicle, covering a zone of about one lane wide on both sides of the vehicle. The zone also extends back approximately 3 m from the rear of the vehicle for an added layer of protection. In this case too, vehicles approaching rapidly from behind cannot be detected within sufficient time.

GM is currently offering its “Side Blind Zone Alert” system in the Buick LaCrosse. Other GM vehicles offering Side Blind Zone Alert for the 2010 model year are the Buick Lucerne; Cadillac STS, DTS, and Escalade ESV; Chevrolet Tahoe and Suburban; and GMC Yukon and Yukon XL.

5.5 “Blind Spot Monitor” from Jaguar

The “Blind Spot Monitor” from Jaguar monitors the blind spot using radar sensors. Based on their installation location and detection range (see Fig. 28.10), these are comparable with the sensors used in the “Blind Spot Information System” from Ford. According to Jaguar, the “Blind Spot Monitor” can take much of the uncertainty out of lane changing. Using radar sensors to remotely cover areas adjacent to the car that cannot be seen either using the mirrors, it is designed to alert the driver to overtaking traffic with an amber warning icon in the external mirrors.

Jaguar is currently offering the “Blind Spot Monitor” in Europe in the XJ and XFR for an additional price of 540 € including VAT. For the Jaguar XF models the customers can purchase the system as part of an equipment package together with other features for 1,170 € including VAT.

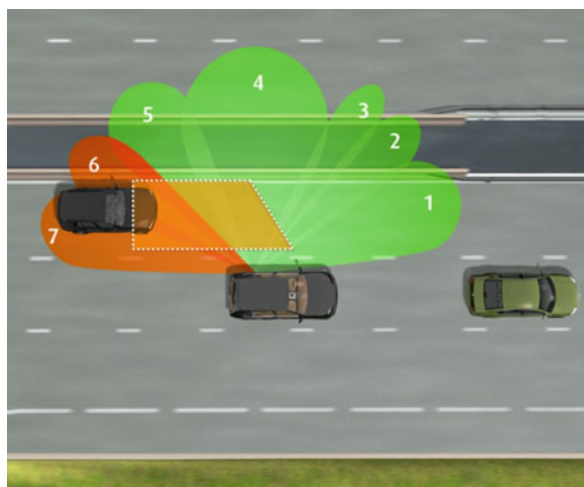
5.6 “Rear Vehicle Monitoring System” from Mazda

Mazda offers its lane change assistant under the name “Rear Vehicle Monitoring System” (RVM). The system operates based on 24 GHz radar and consists of two sensors that are integrated into the corners of the rear bumper, monitoring the area next to and behind the

ego vehicle (see ● [Fig. 28.11b](#)). The system not only informs the driver about vehicles in the blind spot but also about vehicles that are approaching at high speed in adjacent lanes. Two yellow lights fitted under the A-pillar on the left and right warn the driver about a potentially dangerous situation (see ● [Fig. 28.11a](#)). If the driver activates the turn signal in this situation then an acoustic signal also sounds. Mazda uses sensors similar to those employed by Audi, BMW, and VW.

In Europe, Mazda is currently offering the “Rear Vehicle Monitoring System” in the Mazda 3 and Mazda XC7 as standard equipment. In the Mazda 6 models the RVM is part of an equipment package together with other features for 1,390 € including VAT.

A particularly striking feature here is the introduction of the system in the lower vehicle category of the Mazda 3 as standard equipment.



■ Fig. 28.10

“Blind Spot Monitor” from Jaguar. Detection range of a radar sensor (Prova 2007).



a



b

■ Fig. 28.11

“Rear Vehicle Monitoring System” from Mazda (Autobild 2008). (a) The shining light shown in the mirror triangle. (b) Monitored area

5.7 “Blind Spot Assist” from Mercedes-Benz

The “Blind Spot Assist” from Mercedes-Benz also monitors the areas directly alongside and behind the car on the driver’s and front passenger’s sides and warns the driver if it detects a risk of collision when changing lanes. The “Blind Spot Assist” from Mercedes-Benz uses short range radar sensors. These employ ultra wide-band transmission with a mean frequency of 24 GHz, and are integrated on both sides of the rear bumper.

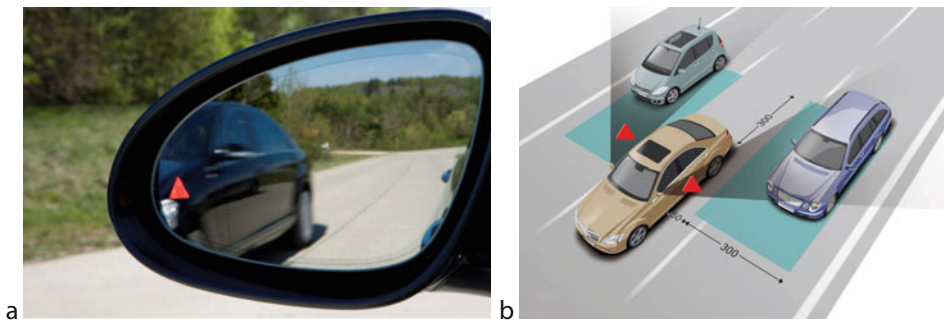
As soon as a vehicle is detected in the monitored area, the driver is informed by illuminating a red warning symbol which is integrated in the exterior mirror glasses. A warning about the threat of a collision is given if a vehicle has been detected in the blind spot monitoring area and the driver has activated the turn signal. This involves a one-off double tone and the red warning symbol flashes. If the turn signal continues to be activated, detected vehicles are permanently indicated by the red light flashing. There is no repetition of the acoustic driver information.

► [Figure 28.12b](#) shows the detection range of the sensors. The monitored area is about 3 m wide, starting at a distance of about 50 cm from the side of the vehicle. The length of the monitored area extends from the driver’s shoulder level to about 3 m behind the rear bumper. This means vehicles approaching at speed from behind cannot be detected within sufficient time. As a result, the customer benefit is restricted, especially on motorways.

An advantageous feature here is the activation speed of the system. Although, at 30 km/h, it is somewhat faster than in the Volvo BLIS, it does nevertheless permit the system to be used in urban areas.

The lights are fitted behind the mirror glass of the exterior mirror, and are almost invisible. These are also used for displaying the system status. If the system is switched on and active then the driver is informed about this by the lamps lighting up yellow.

The radar sensors of the “Blind Spot Assist” must be switched off in certain countries and in the vicinity of radio-astronomical facilities. This is due to the restricted radio certification of 24 GHz radars with wide-band transmission used for automotive applications.



■ Fig. 28.12

“Blind Spot Assist” from Mercedes-Benz (Heise [2007](#)). (a) Red light symbol behind the mirror glass of the exterior mirror. (b) Detection range

The “Blind Spot Assist” was first offered by Mercedes-Benz in the S class in 2007. At present, it is offered as additional equipment in a package together with other radar-based driver assistance systems in the S Class for an additional price of 2558.50 €, in the CL Class for an additional price of 2594.20 € and in the E Class for 2677.50 €, in each case including VAT. The sale price in the R, M, and GL Class is 1,071 € including VAT.

Mercedes recently announced the so called “Active Blind Spot Assist” as an innovation from the latest stage of development: If the driver ignores warnings and the vehicle comes dangerously close to the next lane, “Active Blind Spot Assist” will intervene. Applying braking force to the wheels on the opposite side of the vehicle via the Electronic Stability Program ESP® creates a yaw movement which counteracts the collision course. The system intuitively deactivates as soon as the driver steers against the effects of the braking intervention or the vehicle accelerates.

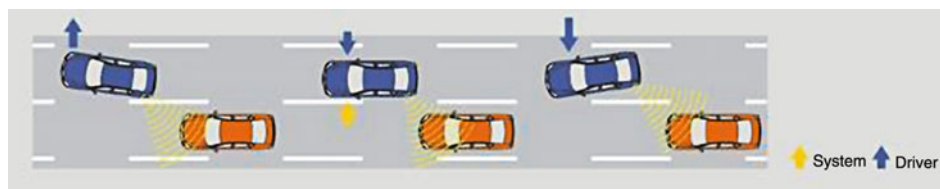
Brake actuation to correct the course occurs between 30 and 200 km/h. The effect is limited to longitudinal and latitudinal deceleration of 2 m/s^2 . When ESP is in OFF mode, “Active Blind Spot Assist” is switched off. Visible warning in the exterior mirror is active up to a speed of 250 km/h (Daimler 2010).

5.8 “Side Collision Prevention” from Nissan/Infiniti

Infiniti has announced the “Side Collision Prevention” (SCP) system developed by Nissan, for its 2011 model year. This is based on two radar sensors fitted at the sides to the rear of the vehicle, which monitor the area to the side adjacent to and behind the ego vehicle.

If a vehicle approaches the ego vehicle from behind, information is provided to the driver in the form of lights that are fitted in the mirror triangle below the A-pillar. If the ego vehicle makes a lane change despite a vehicle approaching from behind then an acoustic signal sounds. In addition, a braking force is applied to individual wheels in order to generate a torque which has the effect of guiding the ego vehicle back into its former lane (see ► Fig. 28.13).

The start of sales of the SCP system has been announced for the 2011 model year of the Infiniti M-Generation. No sales prices have been announced yet.



■ Fig. 28.13

“Side Collision Prevention” from Infiniti (Nissan 2008) Guide-back torque by braking of individual wheels

5.9 “Blind Spot Detector” from Peugeot

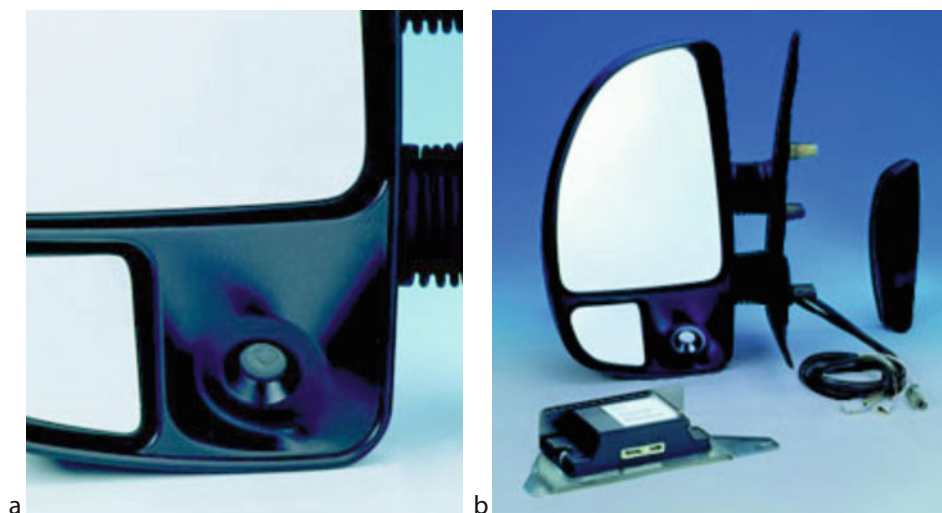
The “Blind Spot Detector” from Peugeot functions similarly to Volvo’s BLIS. This system monitors the blind spot adjacent to the driver’s own vehicle and informs the driver by a light in the A-pillar as soon as a vehicle drives into the blind spot. As with the BLIS, the relevant corridor is monitored by means of a camera integrated in the exterior mirror. In contrast to BLIS, however, only the driver’s side is monitored. Peugeot does not offer any assistance for lane changes on the front passenger’s side. Therefore, the “Blind Spot Detector” does not comply with the system types that are classified in ISO standard 17387.

► [Figure 28.14a](#) shows how the camera is integrated into the exterior mirror. The main system components of the “Blind Spot Detector” such as the control unit and cables are shown in ► [Fig. 28.14b](#).

The system was first offered in the Peugeot Boxer at the start of 2002 at an additional price of 300 € plus VAT. At the same time, the identical system was offered in the Fiat Ducato and Citroën Jumper as well.

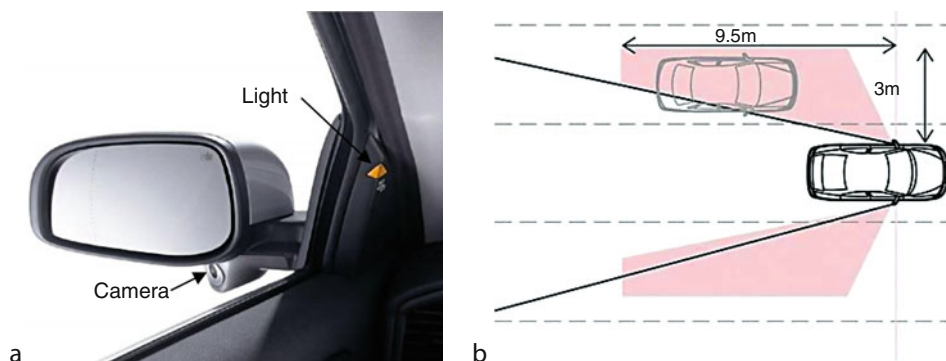
5.10 “Blind Spot Information System” (BLIS) from Volvo

The BLIS from Volvo informs the driver about vehicles that are in the driver’s own blind spot. In particular in dense traffic, this is intended to avoid traffic accidents due to lane changes. The system is based on two digital cameras fitted in the exterior mirrors. These rearward-pointing cameras monitor the traffic in both adjacent lanes to the right and left



■ Fig. 28.14

“Blind Spot Detector” from Peugeot (Ficosa 2002). (a) Integration of the camera in the exterior mirror. (b) Additional system components of the “Blind Spot Detector”



■ Fig. 28.15

Blind Spot Information System (BLIS) from Volvo (Gizmag 2007). (a) Camera integrated in the exterior mirror and yellow light in the A-pillar. (b) Monitoring range of the sensors

of the driver's own vehicle (see ● Fig. 28.15a). If a vehicle enters the blind spot, a lamp in the right or left A-pillar lights up discreetly to inform the driver of this (see ● Fig. 28.15a). BLIS detects all objects that are moving up to 70 km/h faster or 20 km/h slower than the driver's own vehicle.

An advantageous factor here is the low activation speed of the system, since at present it offers effective monitoring of the blind spot at speeds as low as 10 km/h, meaning that the system can also be used in urban areas.

The monitoring range of the cameras is restricted to a corridor about 3 m in width and 9.5 m in length on the left and right adjacent to the driver's own vehicle (see ● Fig. 28.15b). This means vehicles approaching from behind at high speed cannot be detected within sufficient time. This restriction on the customer benefit becomes particularly significant on motorways and when driving at high speeds. There is no escalation of the driver's information, e.g., when the turn signal lever is operated.

This system was first offered by Volvo in the 2005 model year. The additional price for the additional equipment is currently 620 € including VAT, and it is available as an option in all Volvo cars (C30, C70, V50, V70, S40, S60, S80, XC60, XC70, XC90).

5.11 “Side Assist” from VW

VW offers a system that is almost identical to “Audi Side Assist,” marketing it under the name “Side Assist.” Differences between the systems from VW and Audi are apparent in the position of the buttons for activating the system, and the system status indicators. The hardware of the radar sensors and the positioning of the lights in the housing of the exterior mirrors reveal no significant differences between VW and Audi.

VW first fitted “Side Assist” to the Touareg in 2006. The additional price for the system offered as an optional extra is currently 580 € in the Touareg or 585 € in the Phaeton, in

each case including VAT. VW Commercial Vehicles offers the system in the Caravelle, Multivan, California and Transporter, in an equipment package together with other features. The prices for these are between 940 € and 1,390 € plus VAT.

5.12 Summary

- [Table 28.10](#) shows a summary of the lane change assistance systems described in
- [Sect. 5](#), offered by the various vehicle manufacturers.

Only the systems from Audi, BMW, Mazda, and VW offer a “lane change warning” in addition to the “blind spot warning,” which represents an advantage on motorways in particular.

Except for the Peugeot system, all systems do monitor both the driver’s and front passenger’s sides.

All systems (as far as data is available) can be used in urban areas as well due to their low activation speed.

The environment sensors used are either 24 GHz radar sensors invisibly mounted behind the rear bumper or cameras integrated in the exterior mirrors. It is clear that, except for Volvo and Peugeot, all other vehicle manufacturers use 24 GHz radar sensors. At present, no systems using laser sensors are available on the market.

All vehicle manufacturers favor optical driver information with optical indicators in or close to the exterior mirror. The installation location of the optical indicators varies between the A-pillar, the mirror glass of the exterior mirror and the mirror housing of the exterior mirror. Also, the detailed configuration of the optical indicator varies from vehicle manufacturer to vehicle manufacturer, between red and yellow indicators that are configured as lights or illuminated symbols or icons.

The majority of vehicle manufacturers use a two-level driver information function. The detailed configuration of the second information level varies significantly. It ranges from optical, acoustic and tactile information, even as far as interventions in the lateral guidance.

The price for a system that monitors both the driver’s and front passenger’s sides is currently between 495 € and 1,071 €, when offered as additional equipment. Currently, Mazda is the only vehicle manufacturer to offer a lane change assistant as standard.

The ADAC (German motoring organization) conducted a comparative test involving the systems from Volvo, Mercedes-Benz, Audi, and VW at the start of 2008 (ADAC 2008). The test victor was “Audi Side Assist” which, along with VW “Side Assist,” was ranked as “good.”

6 Achieved Performance Capability

The performance capability of the aforementioned lane change assistance has already reached significant levels. However, all of these systems have their limits, and the vehicle manufacturers need to inform drivers of these in the owner’s manual, for example.

Table 28.10

Comparison between lane change assistance systems by the various vehicle manufacturers (✓ yes/X no/– no data)

Vehicle manufacturer	System marketed under name	Blind spot warning	Lane change warning	Driver's and FP's sides	Activation speed in km/h	24 GHz radar in the rear bumper	Cameras in the exterior mirror	Driver information level 1	Driver information level 2	Minimum additional price with/without* VAT
Audi	Audi Side Assist	✓	✓	✓	30	✓	×	Yellow lights in the housing of the exterior mirrors	Light flashes	550 €
BMW	Lane Change Warning	✓	✓	✓	50	✓	×	Red illuminated symbol in the exterior mirror housing	Light flashes, steering wheel vibrates	620 €
Ford	Blind Spot Information System	✓	×	✓	–	✓	×	Yellow lights behind mirror glass	–	495 €
GM	Side Blind Zone Alert	✓	×	✓	–	✓	×	Yellow illuminated symbol behind mirror glass	Light flashes	–
Jaguar	Blind Spot Monitor	✓	×	✓	–	✓	×	Lights behind mirror glass	–	540 €
Mazda	Rear Vehicle Monitoring System	✓	✓	✓	30	✓	×	Yellow lights in the A-pillars	Acoustic signal	Standard equipment
Mercedes-Benz	Blind Spot Assist	✓	×	✓	30	✓	×	Red illuminated symbol behind mirror glass	Light flashes, acoustic signal	1,071 €
Nissan Infiniti	Side Collision Prevention	✓	×	✓	–	✓	×	Yellow lights in the A-pillars	Acoustic signal, vehicle guide-back	–
Peugeot	Blind Spot Detector	✓	×	×	–	×	✓	Yellow lights in the A-pillars	×	300 €*
Volvo	Blind Spot Information system	✓	×	✓	10	×	✓	Yellow lights in the A-pillars	×	620 €
VW	Side Assist	✓	✓	✓	30	✓	×	Yellow lights in the housing of the exterior mirrors	Light flashes	580 €

Almost all vehicle manufacturers are united in pointing out that their system is only an assistant, possibly does not detect all vehicles, and cannot replace attentiveness by the driver. Furthermore, all vehicle manufacturers point out that vehicles may not be detected adequately or, under certain circumstances, may not be detected at all, if the sensors are contaminated or in adverse weather conditions such as rain, snow, or heavy spray.

Furthermore, there may be defects in the lane allocation of other vehicles, because the widths of adjacent lanes cannot be measured but are only estimated. Therefore, it is pointed out that very wide lanes combined with a driving style which places the vehicles at the outer edge of their respective lanes may result in information failing to be provided about vehicles in adjacent lanes. In narrow lanes combined with a driving style that places the vehicles on the inside edge of their respective lanes, there may well be unnecessary instances of information being provided to the driver about vehicles in the next lane but one.

Vehicles with 24 GHz radars installed in their rear bumper cannot use the systems if the sensor apertures are concealed by objects such as bicycle carriers, trailers, or stickers.

In the case of “lane change warning” function it is observed that the driver cannot be informed within sufficient time about vehicles approaching at very high speed. No information is provided to the driver on tight curves with radii less than 200 m.

Therefore, it is apparent that the aforementioned lane change assistance systems could be improved further from a straightforward user’s perspective. The test conducted by ADAC reached the same conclusion [ADAC].

7 Further Developments

The system functions of lane change assistants can be improved further by increasing the performance of environment sensors, for example by increasing sensor ranges and the speed range within which the sensors can be operated reliably.

Furthermore, the system functions can be optimized by using additional sensors. As explained in ● Sect. 6, estimating the lane width can lead either to driver information being provided superfluously, or the information not being provided when it ought to be. The reliability of the lane change assistant can be increased if the system could detect the position of the adjacent lane as well as the position of the target vehicle.

Different methods can be used for this, depending on the sensor platform in the vehicle: If the lane behind the vehicle is directly detected by rearward-looking sensors with an appropriate range, this information can be used directly or after filtering.

If the range of the sensors to the rear is restricted or if the lanes are detected by forward-looking sensors, then they can continue to be tracked for a certain length of time by means of odometry estimation even outside the area that can be scanned by the sensors. Such information can be obtained, for example, from the data of yaw rate and wheel speed sensors which form the basis for the electronic stability program (ESP). A certain relative accuracy is required in estimating the position of the ego vehicle, target object and

adjacent lanes, because all error variances have an influence on the precision of lane allocation.

This indirect estimation is particularly appealing if it is possible to rely on the environment sensors of other driver assistance systems. For example, a forward-pointed camera is often installed in the ego vehicle for applications such as lane departure warning or lane keeping support.

At present, lane change assistants only help the driver to decide whether a lane change is possible or not. The lane change itself must be performed independently by the driver, however. This step could also be assisted with the help of sensors detecting lane markings, in combination with an electronically controllable steering actuator system; suitable steering torques would enable lateral guidance to be assisted during a lane departure or lane entry maneuver, or indeed even automated.

Acknowledgments

The authors would like to express their gratitude at this point to Prof. Winner of Darmstadt Technical University for his assistance in structuring and designing this article, as well as for his initiative of creating the “Driver Assistance Systems Handbook” (Winner et al. 2009).

References

- Accident database of Volkswagen Accident Research and GIDAS (German In-Depth Accident Study)
- ADAC homepage (2008) Test Spurwechsel-Assistenten. http://www.adac.de/_mm/pdf/Test%20Spurwechsel%20Assistenten_23685.pdf. Accessed 7 Oct 2010
- Article in Autobild.de (2008) http://www.autobild.de/artikel/mazda6-2.2-mzr-cd_806923.html. Accessed 25 May 2010
- Article in Gizmag Internet magazine (2007) Volvo launches blind spot information system (BLIS). <http://www.gizmag.com/go/2937/>. Accessed 11 Sept 2008
- Article on Heise Online (2007) Mercedes: new blind spot Assistant. <http://www.heise.de/autos/S-und-CL-Klasse-Neuer-Totwinkel-Assistent-fuer-mehr-Sicherheit-beim-Spurwechsel-/artikel/s/4517>. Accessed 11 Sept 2008
- Daimler global media site (2010) New driver assistance systems: premiere: active blind spot assist and active lane keeping assist with brake actuation. <http://media.daimler.com/dcmedia/0-921-658892-1-1298797-1-0-0-1299004-0-0-11702-0-1-0-0-0-0-0.html>. Accessed 29 June 2010
- Ehmanns D, Aulbach J, Strobel T, Mayser C, Kopf M, Discher C, Fischer J, Oszwald F, Orecher S (2008) BMW 7 series: active safety and driver assistance. ATZextra:114–119
- Ford press release on its Blind spot information system (2009) <http://media.ford.com/images/10031/BLIS.pdf>. Accessed 25 May 2010
- GM homepage (2010) Side blind zone alert in Buick LaCrosse can help avoid lane change mishaps. http://www.gm.com/corporate/responsibility/safety/news/2010/blind_spot_040910.jsp. Accessed 25 May 2010
- Homepage of FICOSA (2002). <http://www.ficosa.com/home.php>. Accessed 26 Feb 2002
- ISO standard 17387 (2008) Lane change decision aid system

- Nissan homepage (2008) “2008 advance technology briefing” (First part). <http://www.nissan-global.com/EN/IR/INSIDE/-INSIDE-SP/ATB2008/page02.html>. Accessed 26 May 2010
- Owner’s manuals of the Audi Q7 and VW Touareg (2008)
- Popken M (2006) Audi side assist. Hanser Automotive electronics + systems:54–56, 7–8
- Prova gallery (2007) Jaguar blind spot monitoring: the car looks back. <http://www.prova.de/archiv/2007/00-artikel/0141-jaguar-toter-winkel/index.shtml>. Accessed 25 May 2010
- 7-Series Forum (2008) Press photos published by BMW of the new 7-Series. http://www.7-forum.com/modelle/f01/galerie_730d_750li.php. Accessed 25 May 2010
- Winner H, Hakuli S, Wolf G (2009) Handbuch Fahrerassistenzsysteme, Vieweg + Teubner Verlag, ISBN 978-3-8348-0287-3
- Vukotich A, Popken M, Rosenow A, Lübcke M (2008) Driver assistance systems (special edition) ATZ and MTZ:170–173, 2

29 Steering and Evasion Assist

Thao Dang¹ · Jens Desens¹ · Uwe Franke¹ · Darius Gavrila² ·
Lorenz Schäfers¹ · Walter Ziegler¹

¹Group Research and Advanced Engineering, Driver Assistance and
Chassis Systems, Daimler AG, Sindelfingen, Germany

²Group Research and Advanced Engineering, Driver Assistance and
Chassis Systems, Daimler AG, Ulm, Germany

1	<i>Introduction</i>	760
2	<i>Potential of Evasion Versus Braking</i>	761
3	<i>System Layout</i>	763
3.1	Overview	763
3.2	Requirements for Situation Analysis	764
3.3	Requirements for Environment Perception	765
3.4	Actuators	768
3.4.1	Steer Torque Actuator	768
3.4.2	Steer Angle Actuator	769
3.4.3	Rear Wheel Steering	769
3.4.4	Warping the Suspension	769
3.4.5	Single-Sided Braking	769
3.4.6	Torque Vectoring	769
3.4.7	Discussion	770
3.5	HMI and Customer Acceptance	770
4	<i>Case Studies</i>	771
4.1	Survey of Research Activities of Industry and Academia	772
4.2	The Daimler Automatic Evasion Assistance for Pedestrian Protection	773
4.2.1	Motivation	773
4.2.2	System Description	774
4.2.3	Experiments	778
5	<i>Conclusion</i>	780

Abstract: Steering and evasion assistance defines a new and future class of driver assistance systems to avoid an impending collision with other traffic participants. Dynamic and kinematic considerations reveal that an evasive steering maneuver has high potential for collision avoidance in many driving situations. Three different system layouts are described: driver-initiated evasion, corrective evasion, and automatic evasion assistance. Since an automatic steering intervention is a challenging and responsible task, the technological requirements for situation analysis and environment perception are stated. Many technical solutions for a steering intervention are conceivable; therefore several actuator concepts are discussed and assessed with respect to human machine interface (HMI) impacts. A short survey of research activities of industry and academia is given. As an example for a research level prototype, the Daimler automatic evasion assistance system for pedestrian protection is presented in detail. Based on binocular stereo vision, crossing pedestrians are detected by fusion of a pedestrian classification module with a 6D-Vision moving object detection module. Time-To-X criticality measures are used for situation analysis and prediction as well as for maneuver decision. Tested on a proving ground, the prototype system is able to decide within a fraction of a second whether to perform automatic braking or evasive steering, at vehicle speeds of urban traffic environment. By this it is shown that automatic steering and evasion assistance comes to reality and will be introduced stepwise to the market.

1 Introduction

Driver assistance systems available today support the driver during normal driving as well as in critical situations by warnings and – in case of an impending collision – by partial or full braking. Emergency braking systems are able to mitigate or even prevent a collision in many cases, but there are situations left where an obstacle appears suddenly, e.g., a pedestrian is crossing the road and even full braking is too late and will not avoid the collision. In these cases a steering intervention is an additional option to prevent the collision. Steering intervention has to be apprehended as a new and second path to resolve impending collision situations and therefore defines a new class of driver assistance systems. Up to now no product level system uses of a steering intervention to avoid a collision. This is caused by the complexity, the required challenging technologies, and the product liability aspect of such an intervention. However during the last 5–10 years several research activities of automotive manufacturers, suppliers, as well as universities could be observed. This chapter will give an overview about the safety potential of such a system (➤ Sect. 2), the technological requirements (➤ Sect. 3), as well as research activities as an example for future evasion assistance systems (➤ Sect. 4).

The safety potential of a steering and evasion assist depends on the situation and mainly on the relative velocity between the system car and the obstacle. ➤ Section 2 will point out the collision avoidance potential compared to systems with exclusive braking intervention.

What kind of system will a steering and evasion assist be? With the above given system description three different system characteristics are conceivable:

- *Driver-initiated evasion assistance:* The driver has to do the first step. If there is an obstacle in front and the driver indicates by his or her steering activity that he or she wants to circumnavigate the obstacle, the system will help him or her to perform the maneuver stable and safely.
- *Corrective evasion assistance:* If braking will not be sufficient to prevent a collision, but the amount of space needed for the evasion maneuver is somewhat small, e.g., half of the width of the car, the steering action is initiated by the system itself.
- *Automatic evasion assistance:* The steering action is initiated by the system itself, if braking will not be sufficient to prevent a collision. The amount of space needed by the system for the evasion maneuver is in principle not limited but depends on the situation.

These different system characteristics will be described in [Sect. 3](#), where an explanation is given about the requirements for environment perception, the situation analysis, the applicable actuators, and the HMI and customer acceptance.

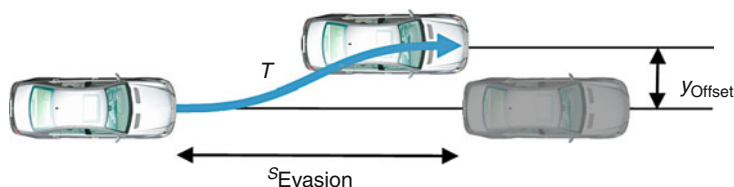
2 Potential of Evasion Versus Braking

In the following the safety potential of an evasion maneuver will be derived from dynamic and kinematic considerations.

Advanced driver assistance systems (ADAS) use sensors to observe the environment. Based on the environment data, an active safety system decides whether there is an imminent risk of collision and whether an automatic maneuver has to be performed to avoid or mitigate the accident. The driver remains responsible for the driving task also in presence of assistance systems. Therefore a design principle of active safety systems is to intervene not until the physically latest possible point in time. This point in time is defined by the driver's ability to avoid a collision by braking, steering, or a combination of both, or – in some cases – by accelerating.

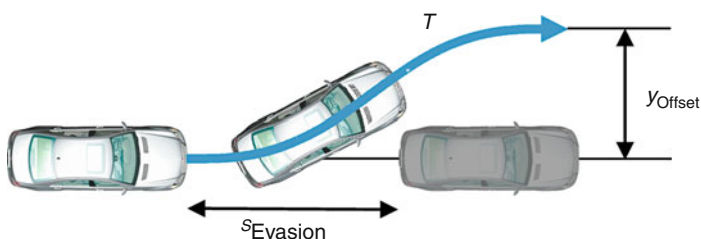
To evaluate the potential of evasion vs. braking independently from a specific assistance system, the steering distance to an object ahead will be compared with the braking distance. The latter is easily computed on basis of the relative velocity between the vehicle and the object and the full braking deceleration of the vehicle in case of an emergency braking maneuver.

The steering distance S_{Evasion} at a given relative velocity depends on the driven maneuver trajectory T , the desired lateral offset γ_{Offset} , and a predefined maximum lateral acceleration $\alpha_{\gamma, \text{max}}$. The lateral offset at least needs to reach the sum of half the vehicle width, half the object width, and a safety margin. The steering distance then is the distance needed to reach the lateral offset when driving trajectory T , see [Fig. 29.1](#). Shorter steering distances can be achieved by changing the respective parameters.



■ Fig. 29.1

Steering distance and lateral offset during an evasive maneuver



■ Fig. 29.2

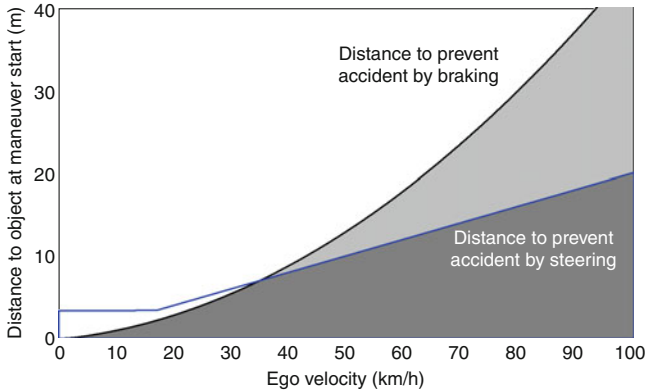
Changing the evasive maneuver to a shorter steering distance yields a larger lateral offset

Typical trajectories for evasive maneuvers include clothoid and sigmoid functions, the latter yielding shorter steering distances than the first. The lateral acceleration is in principle only constrained by the physical limits, i.e., the current road friction coefficients, the wheel characteristics, etc. Increasing the maximum lateral acceleration leads to a shorter steering distance as well. Finally, the lateral offset can be increased, passing the object during the evasive maneuver – and not only reaching the end – thus also reducing the steering distance. Of course, this is only possible if other traffic and boundaries allow realizing this increased offset (► Fig. 29.2).

The potential of evasion arises in situations where the braking distance exceeds the steering distance, i.e., steering potentially still avoids a collision where braking does not. Such traffic situations comprise collisions with standstill objects or slowly moving vehicles and high relative velocity as well as accidents with perpendicular slow moving traffic participants (e.g., pedestrians, cyclists) where frequently a small lateral offset is sufficient to avoid the collision. In addition, if the friction coefficient of the road decreases, e.g., on a wet road, the benefit of evasion vs. braking expands to lower relative velocity.

► Figure 29.3 compares braking distance and steering distance for a sigmoid trajectory at different velocities on a dry road. The vehicle reaches a lateral offset of 1 m at the inflection point of the evasive path and a total lateral offset of 2 m. In this example, the braking deceleration is assumed to be -10 m/s^2 and the maximum lateral acceleration is bounded to 6 m/s^2 .

Starting from about 35 km/h an evasive steering maneuver enables collision avoidance at lower distances to the object than full braking with increasing benefit at higher speeds.



■ Fig. 29.3

Comparing braking distance and steering distance at different velocities

3 System Layout

3.1 Overview

Driver assistance systems can be classified according to the degree of automation. Regarding steering and evasion assist, three system concepts with increasing complexity have to be deliberated:

- Driver-initiated evasion assistance
- Corrective evasion assistance
- Automatic evasion assistance

A *driver-initiated evasion assistance system* supports in critical driving situations, in which the driver starts steering to drive around an obstacle in front. Such systems comprise at least one sensor like radar, lidar, or camera that detects obstacles in the driving corridor ahead. When an object is detected and the situation analysis predicts a collision, system activation requires the driver steers in a specific, speed-dependent distance section ahead of the object. Typically, this section will be closer than distances used for overtaking, hereby distinguishing a critical driving situation from a normal one. In those situations the system shall improve the driver's steering operation and vehicle maneuverability. This can be achieved by using a steering actuator which creates a torque that assists and guides the driver on a predetermined trajectory around the object. If available, an active suspension system like active dampers or active roll stabilization can additionally improve the handling performance of the vehicle. The system can be designed in such a way that the driver is always able to overrule the system intervention. Since the driver initiates the steering maneuver, the evasion assist system does not necessarily need to check for other traffic participants or obstacles in the evasive trajectory as this has been accomplished already by the driver.

A *corrective evasion assistance system* targets potential collision scenarios, where already a small lateral offset, e.g., half the width of a vehicle, helps to prevent an accident. Such scenarios vary from less critical situations like cycles or motorcycles driving ahead, parked or broken down vehicles, or other objects at the roadside extending into the traffic lane of the own vehicle up to dangerous scenarios where objects (e.g., pedestrians) are crossing the lane. Again, an appropriate sensor set perceives the object; the distance to the object has fallen below the braking distance, and a situation analysis module (cf. ● Sect. 3.2) predicts an impending collision. Then the assistance system automatically triggers an evasive maneuver to avoid the collision. In many cases it will be beneficial if the system combines evasion with braking. As the system starts the maneuver automatically, if the driver fails to react timely, a good knowledge about the traffic environment is needed. The system should ensure not to collide with other traffic participants nor objects during the evasive maneuver.

An *automatic evasion assistance system* is capable of coping with many different traffic situations. Here, the lateral offset and hence the amount of space needed to drive the maneuver is in principle not limited but depends on the demands of the situation. Such a system requires a widespread and detailed knowledge of the environment all around the ego vehicle. As before, when the system foresees an imminent collision and the driver neither brakes nor steers at the latest possible point in time, respectively, the system will perform an automatic evasive maneuver.

The requirements for such systems are described in the following sections.

3.2 Requirements for Situation Analysis

The objective of situation analysis in the context of advanced driver assistance systems (ADAS) is to understand and analyze a given traffic situation. Such knowledge can be exploited to derive automatic actions of the vehicle. Thus, situation analysis is closely related to the cognitive capabilities of intelligent vehicles (e.g., Stiller et al. 2007). Generally, situation analysis of an ADAS includes: *modeling* of the vehicle's environment (what is known about the current scenario?), *classification* of driving situations (what kind of traffic situation the system is confronted with? what are the current maneuvers of the traffic participants?), *prediction* (what are possible actions of all objects including the ego vehicle?), and *criticality assessment* (how severe are the results of these actions).

ADAS that mitigate collisions in longitudinal traffic by automatic braking have been studied extensively and commercial solutions are available on the market (e.g., the Mercedes-Benz PRE-SAFE® Brake, Honda's Collision Mitigation Brake System, Toyota's Pre-Collision System, and others). Typically, the situation analysis layer of such systems is relatively simple and can be sketched as follows: For ADAS in highly structured scenarios, the environment model consists mainly of other vehicles with position and speed information. These vehicles are classified as relevant objects by associating them to the ego vehicle's driving path and imposing thresholds on the object's confidence measures as provided by the sensors. If a collision with a relevant object is detected, the criticality of the

current situation can be evaluated by the time-to-react (TTR, cf. Hillenbrand et al. 2006). The TTR is the remaining time for a human driver to avoid an imminent collision by emergency braking with full deceleration, steering with maximum lateral acceleration, or a kickdown maneuver. Thus, it can be computed as the maximum of the time-to-brake (TTB), time-to-steer (TTS), and time-to-kickdown (TTK).

For evasive steering, however, the requirements for modeling, prediction, and criticality assessment are significantly more challenging. Since changing the vehicle's course could lead to a collision with another traffic participant, a reliable understanding of the vehicle's environment is of paramount importance. The environment model should not only include objects in the front and relevant side. It should also incorporate information about trafficable road (limited by curbs, lane markings, etc.).

The accurate prediction of the trajectories of all objects in the vehicle's environment imposes high requirements on the sensors (cf. [Sect. 3.3](#)). If the evasive maneuver exceeds the own lane, accurate measurements are needed, e.g., of the velocity of an oncoming vehicle in an adjacent lane or the motion of a crossing pedestrian. To predict the possible emergency actions of the ego vehicle, the system has to plan a safe route that will not result in collision with obstacles. The generated trajectory has to be feasible with respect to the vehicle's dynamics. In robotics, such tasks are often referred to as non-holonomic motion planning problems with dynamic obstacles. A variety of solutions has been proposed in the literature (LaValle 1998; Fiorini and Shiller 1996), yet the computational complexity of many of the proposed algorithms prohibits the application on current automotive hardware. To overcome this limitation, efficient planning algorithms to evaluate possible avoidance maneuvers in highly structured scenarios have been introduced (Schmidt et al. 2006). The DARPA Urban Challenge gave rise to several interesting approaches on trajectory generation (Ziegler and Stiller 2009; Werling et al. 2010) that may become feasible for realization in ADAS.

Current collision mitigation systems by braking usually consider only one single object in our lane. Decisions for evasive steering, however, require a criticality assessment that considers multiple objects. In addition, situation analysis has to ensure a reliable decision making even in presence of reasonable sensor and prediction uncertainties. Recent work on handling uncertainty in situation analysis and on the theory of hybrid reachable sets may prove beneficial to accomplish this task (Althoff et al. 2009; Schubert et al. 2010). To ensure a collision-free automatic evasive maneuver, situation analysis has to be closely coupled with vehicle control. Ideally, the prediction of the ego vehicle's behavior will account for the controller characteristics as described in [Sect. 4](#).

3.3 Requirements for Environment Perception

As stated in the preceding section, evasion poses significantly higher demands on the environment perception than pure braking. Roughly speaking, in a certain range depending on the current speed and the possible evasion trajectories relevant moving objects should be detected and classified, their motion state and size should be

determined, and the free space limited by static obstacles should be measured by means of a proper set of sensors. Assuming a maximum speed of 20 m/s (i.e., 72 km/h) of both, the ego vehicle and an oncoming car, and a time of 1 s until the laterally intruding obstacle is passed and another second to steer back or to come to a full stop, the required look-ahead distance is approximately 80 m. This may act as a rule of thumb for the following discussions. If the task is restricted to slower urban traffic and concentrates on vulnerable road users, in particular pedestrians, the requirements are less demanding.

One could argue that it is sufficient to survey the area in front of the car only. In case of an emergency, most human drivers will not check the areas beside the car but will react immediately. This works usually fine since the probability that someone is overtaking exactly that moment when an evasion situation occurs is extremely small. However, an active safety system should take into consideration that possible evasion trajectories could intersect with the driving path of currently overtaking cars or motorcycles. Fortunately, there are several types of so-called blind spot monitoring systems (based on vision as well as radar) on the market that could be used for this task. If the side area is not checked, the risk of an unwanted lateral collision can be reduced if the evasion trajectory is optimized for minimum lateral deviation. The knowledge of the lateral position of existing lane markings that define “my” lane may also be useful in this optimization step.

As long as the system is designed for urban traffic, there seems to be no need to additionally monitor the area behind the car. Since steering is an option for higher velocities only, relative speeds are small and the risk to endanger approaching traffic participants seems to be negligible. This may be different for highway situations which are out of the scope of this chapter.

As the surveillance of the area besides and behind the car is a well-understood problem, the system can concentrate on the area in front of the car. For the sketched tasks, several active as well as passive sensors could be used:

- *Radar*: Recently developed automotive long range radar sensors have beams with a horizontal angle of less than 1° , a field of view of more than 20° , and operating distances which are large compared to the requirement in the urban as well as rural scenario. Therefore they are optimum for the detection of oncoming traffic objects. However, they are still not the preferred sensor for static as well as crossing objects, in particular pedestrians that have a relatively small radar cross section.
- *Lidar*: Laser scanners became famous in 2007, when most finalists of DARPA's Urban Challenge based their autonomous cars on a high-end sensor developed by [Velodyne](#). This sensor offers 64 scan lines with 4,000 measurements per turn, 10 turns per second. The range is about 70 m, the precision is within centimeters. In contrast to radar systems, speed cannot be measured but only estimated by tracking of objects. However, at the time being no scanner fulfills the requirements “performance” and “price” at the same time.
- *Camera*: Vision has become a powerful and cheap solution for driver assistance. Lane Departure Warning and Traffic Sign Recognition are well-established systems, but also

obstacle detection based on a combination of stereo and classification is commercially available. Camera systems are not restricted to the visible range; in fact far infrared sensors are used to detect animals and pedestrians especially at nighttime. Thanks to the high spatial and temporal resolution, cameras will become increasingly important for advanced driver assistance. A large research community working on sophisticated computer vision algorithms is constantly pushing the limits. The problem of camera-based systems is their sensitivity with respect to adverse weather and lightning conditions.

These three types of sensors can operate autonomously, independent from infrastructure. Of course, environment perception as well as situation analysis can be supported by map information as well as vehicle-to-vehicle communication. The latter will help to detect other vehicles earlier, especially if they are hidden by other objects.

It is evident that the performance of the aspired safety system will highly depend on the reliability and accuracy of the sensing system and the time it needs to detect potentially dangerous situations. As a matter of course an automatic evasion system will try to maximize availability and reliability by proper sensor fusion. The current trends in sensors for driver assistance indicate that a fusion of radar and vision is the most promising combination.

In the following, the requirements shall be reconsidered in more detail. ♦ [Figure 29.4](#) shows an urban situation with a pedestrian coming from the right side. The parking car partially hides the pedestrian, while an oncoming vehicle blocks the space necessary for an evasion maneuver. A correct interpretation of this situation requires that:

1. The oncoming car is detected, which is trivial. A vision-based algorithm (Barth and Franke 2009) was published that allows estimating the complete motion state including the yaw rate of moving vehicles.
2. The endangered pedestrian is detected and his or her motion (i.e., velocity and acceleration) is estimated, even if he or she is partially hidden (Enzweiler et al. 2010).



■ Fig. 29.4

One second before the collision. The camera system has estimated the motion state of the oncoming car, and, at the same time, has detected the crossing pedestrian

It would be highly advantageous if this task could be solved *before* he or she steps on the road, since our car can drive at a speed of 0.6 m/frame and some frames delay can make a significant difference

3. If there would be no oncoming traffic, the available free space would be limited by the trees, the parking cars behind the intersection and the curb. While the trees are well-visible objects for a camera system, the detection of the curb is more challenging (Siegemund et al. 2010).
4. Additionally, it would be highly desirable to derive some hints on the pedestrian's intention. Is the pedestrian going to stop – or will he or she go? This is a new question and research has just started.

Binocular stereo vision has the potential to generate a precise three-dimensional model of the current situation (Gehrig and Franke 2007), to detect independently moving objects in minimum time (Franke et al. 2005) and to classify pedestrians even if they are partially hidden. ➤ Section 4.2 will show the state-of-the art in stereo vision.

It is worth to mention that the situation analysis (and the function itself) does not only ask for a comprehensive detection scheme, but also for the confidence of the sensing system regarding the delivered data. This is an additional challenging requirement to be solved within future work.

3.4 Actuators

There are several technical possibilities to influence the lateral movement of the vehicle. Depending on the purpose of the assistance system they are more or less suitable. The following section describes the technical solutions, their advantages, and disadvantages.

3.4.1 Steer Torque Actuator

The actuator adds a steer torque to the torque which the driver applies via the steering wheel.

The driver can suppress the intervention right from the beginning and he or she is given a very intuitive feedback via steer torque and steer angle. The intervention normally is already technically secured by the electric power steering. The steer torque actuator is not appropriate for quick interventions with higher torque because of the risk of injury of the driver's hands. So the steer torque actuator is suitable only for less dynamic interventions but even over a longer time.

All the other actuators share the following advantage and disadvantage. They are appropriate for quick interventions because they are not turning the steering wheel and therefore cause no risk of injury of the driver's hands. But this also means that the driver may be irritated during longer interventions because there is no correlation of the lateral vehicle movement with the steering wheel angle. They also share the danger that

a backlash of the driver on sudden interventions may lead to the wrong direction. This makes them suitable only for short interventions but even with high dynamic.

3.4.2 Steer Angle Actuator

This actuator adds a steer angle to the angle which the driver applies via the steering wheel. If there is no additional torque actuator the driver has to hold a small reaction torque if intervention should effect the vehicle movement and therefore he or she gets a haptic wrong feedback because reaction torque is contrary to the wanted vehicle reaction. The steer angle actuator is suitable only for short interventions but even with high dynamic.

3.4.3 Rear Wheel Steering

The rear wheels can be steered independently from the driver-steered front wheels. Rear wheel steering is suitable only for short interventions but even with high dynamic.

3.4.4 Warping the Suspension

If the wheel load is shifted from the left to the right at one axle and vice versa at the other axle this causes side forces which induce a yaw rate without causing a rolling movement of the vehicle body. This can be done by active suspension systems. But it has only a limited influence on the lateral movement (around $2^\circ/\text{s}$ yaw rate) of the vehicle. Warping the suspension is suitable only for small and short interventions but even with high dynamic.

3.4.5 Single-Sided Braking

Braking the wheels only at one side of the vehicle causes a yaw rate. This can be done by ESP systems. This intervention normally is already technically secured by ESP. Problems are that any intervention also causes a deceleration and could be used only for rare interventions because it causes wear of the brake pads. Therefore single-sided braking is suitable only for rare and short interventions but even with high dynamic.

3.4.6 Torque Vectoring

Unequal drive torque between left and right wheels is causing a yaw rate, which can be realized by an active differential. This is only possible in situations when a positive drive torque is applied or with a wheel individual drive concept, e.g., with electric engines. Torque vectoring is suitable only for short interventions but even with high dynamic.

3.4.7 Discussion

Which solution is best differs widely with the purpose of the assistance system. A system which compensates disturbances like side wind or lateral slope will use other actuators than a system which wants to influence the trajectory of the vehicle to avoid a collision. If the purpose of the assistance system is an evasive maneuver two cases have to be distinguished:

1. The system's intention is to support the driver's steering action to avoid a collision. In this case it has to give the driver an intuitive advice to steer. This is done best directly at the steering wheel and therefore the steer torque actuator is recommended.
2. The system's intention is to avoid a collision by an automatically initiated evasion maneuver (with corrective or large lateral offset as mentioned in [Sect. 3.1](#)). In this case the following actuators are suitable.
 - Steer torque actuator with limited torque
 - Steer angle actuator
 - Rear wheel steering
 - Single-sided brakingTheir actions are quick and strong enough to change the trajectory of the vehicle significantly.

3.5 HMI and Customer Acceptance

The HMI design has to consider the specific situation of an evasive steering maneuver as well as the system layout. As mentioned in [Sect. 3.1](#) the application scenarios vary from less critical situations, where only a light intervention is sufficient to resolve the situation, up to dangerous situations, where a sudden and unexpected collision with a crossing object, e.g., a pedestrian is imminent. In the latter case time is up for warning and even for braking and the HMI design has to concentrate basically on the modality of the steering intervention. Acoustical and/or optical warnings are reasonable if there is enough time for the driver to react. However, in case of sudden and unexpected collision a spontaneous intervention with minimal delay is essential. The intervention may be accompanied by acoustical or optical warnings, but this is not of decisive importance.

The goals of an efficient and ergonomic steering intervention are:

- To perform the evasion maneuver fast and stable
- To give the driver a good understanding “what's going on here”
- To give the driver a chance to overrule the intervention and overtake the responsibility as fast as possible
- Neither to irritate the driver nor to provoke wrong reactions

In [Sect. 3.4](#) all kinds of appropriate actuators have been discussed and assessed. The specific HMI design depends on the functional concept and layout of the system.

Concerning *driver-initiated evasion assistance*, a very tight interaction between driver and system is necessary. The system intervention only has to support or complement driver's action and therefore the steer torque actuator will be the best choice to give the driver a direct feedback.

Concerning *corrective or automatic evasion assistance*, HMI design has to differ between light and strong interventions:

- Light interventions will not dramatically change the vehicle state and driving situation and therefore the driver must not directly feel the intervention “in his or her hands.” As a consequence all proposed actuators are applicable (steer torque actuator, steer angle actuator, rear wheel steering, single-sided braking).
- Strong interventions provoke a significant change in driving situation. To safeguard an adequate driver reaction, he or she should clearly know what happens and understand where the evasive maneuver does come from. As motion and torque of the steering wheel immediately communicate the driver what's going on, the steer torque actuator seems to be the best choice. On the other hand the torque has to be limited due to the risk of loss of controllability as well as the risk of injury of driver's hands as mentioned in [Sect. 3.4](#). Therefore, if the automatic motion of the steering wheel exceeds certain values (see next paragraph), the steering intervention should be supported by rear wheel steering or single-sided braking or should exclusively be realized with a steer angle actuator.

Strong steering intervention by additional steering torque has to consider several limits due to controllability and acceptance reasons. The most important parameters are steering wheel angle, velocity, and acceleration as well as the additional steering wheel torque itself. Basic studies investigated the interrelationship between those parameters and human behavior in terms of steering quality and controllability, e.g., Neukum (2010). As controllability has to be recovered after the maneuver, the design of the evasion trajectory itself has to take care for an easy handing over when the maneuver is completed: It has to be limited in short duration (e.g., <1 s), the vehicle course should be stable at any time, and the yaw angle and yaw velocity of the vehicle should be zero when it is finished. Anyway, controllability and customer acceptance have to be approved by real driving tests with a sufficient number of test persons.

4 Case Studies

Since the 1980s, several research programs on autonomous vehicles have been conducted, finally leading to the DARPA Urban Challenge in 2007. Numerous publications on path control for automated vehicle guidance, active steering systems, steering controllers, and the like have been released. A focused view on evasive steering support in research or production cars is given here.

4.1 Survey of Research Activities of Industry and Academia

From 2002 to 2006, Darmstadt University of Technology and Continental Automotive Systems conducted the PRORETA project which investigated the collision avoidance potential of emergency braking and emergency steering in case of stand-still objects or objects cutting into the line in front of the own vehicle (Isermann et al. 2008). A demonstrator vehicle (Volkswagen Golf) was equipped with an electro-hydraulic brake and an active front steering and publicly demonstrated at the end of the project.

The system detects objects by a fusion of scanning laser and video. In case of a threatening collision and depending on the traffic situation, the system elects one of three possible intervention schemes: braking, steering, or a combination of both. In case of a suddenly appearing obstacle or unexpectedly blocked lane an automatic emergency evasion maneuver is conducted, if possible. Based on the information from the environmental sensors, the necessary evasive trajectory is calculated. A lateral controller then automatically guides the vehicle round the obstacle on the predefined evasive path. Different controllers were implemented and tested.

In February 2006, Toyota presented the Lexus LS 460 at the Geneva Motor Show, equipped with microwave radar and a stereo camera. The technical features include an emergency steering assist (Suzumura et al. 2007). When the system detects a possible collision with an object ahead, emergency steering assist enables the car to react more spontaneously on the driver's steering commands and thus improving evasive maneuvering initiated by the driver. For this purpose, variable-gear-ratio steering, vehicle-dynamics integrated management, and adaptive variable suspension are combined resulting in a more direct steering gear ratio, a selective use of the brakes, and a stiffer chassis suspension.

Recently Bosch and Continental independently published two similar approaches of an emergency steering assist, both based on the concept of driver-initiated assistance.

The Bosch system is called evasive steering support (Fausten 2010). It uses microwave radar to detect an obstacle in front. If there is a risk of a rear-end collision and the driver starts to steer, the system will support him or her to follow an optimal evasion trajectory according to the following support strategy: As long as the driver steers on the optimal trajectory, there is no support. If the driver overreacts, there is a corrective torque on the steering wheel. When the driver underreacts, the system supports the driver with an additional torque on the steering wheel. The system limits the steering torques and thus guarantees the controllability by the driver at any time.

The Continental emergency steer assist combines an environmental sensing with situation-dependent adaptation of electrically controllable chassis components such as electric power steering, electronic stability control, and optional rear wheel steering system (Hartmann et al. 2009). As before, microwave radar detects a leading object. If a potential collision is foresighted, the system prepares the vehicle for an optimal driving

stability by activating specific modes of the electronic stability control and the rear wheel steering control. Already small instabilities of the vehicle are compensated by early and well-directed damping of vehicle overshoot reactions. Upon the driver starting to steer, the system supports on a maneuvering level to keep the vehicle on an optimal evasion trajectory by either an additional steering torque or torque vectoring. The system is designed in such a way, that the driver can overrule it at any time.

4.2 The Daimler Automatic Evasion Assistance for Pedestrian Protection


4.2.1 Motivation

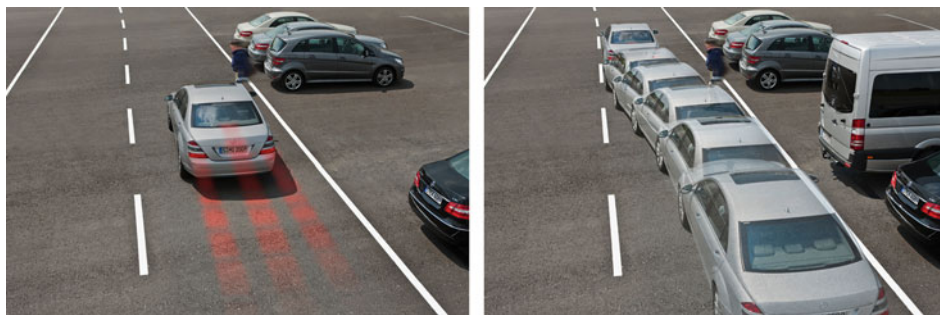
The most vulnerable traffic participants are arguably the pedestrians; about 5,700, 4,700, and 2,300 pedestrians are killed yearly in traffic in the EU, USA, and Japan, respectively (IRTAD 2006). These figures correspond to approximately 18%, 11%, and 32% of all traffic fatalities in the respective regions.

Traditionally, pedestrian protection has been approached from a passive safety perspective. This has involved vehicle structures (e.g., bonnet, airbags) which expand during collision in order to minimize the impact of the pedestrian body hitting the vehicle. Although important, passive safety measures are limited in their ability to reduce collision energy because of the very short time span between initial bumper contact and the impact of the pedestrian on the bonnet or windshield. Moreover, passive measures cannot account for injuries sustained in the secondary impact of the pedestrian hitting the pavement.

There is a lot of interest, therefore, in the development of active driver assistance systems, which use sensors to search the vehicle surroundings for pedestrians. They can detect dangerous situations ahead of time, and warn the drivers or even automatically control the vehicle. Such systems are particularly valuable when the driver is inattentive (e.g., programming the navigation unit, or head turned to the back seat).

Gandhi and Trivedi (2007) provide a general survey on passive and active pedestrian protection methods, discussing multiple sensor modalities (e.g., cameras in visible/NIR/FIR spectrum, radars, laser range finders) and methods for collision risk assessment. Enzweiler and Gavrila (2009) focus in a more recent survey on techniques for video-based pedestrian sensing. A large image dataset is made publicly available for benchmarking purposes.

In this section, a recent research prototype system for active pedestrian safety is discussed, developed at Daimler R&D, which combines sensing, situation analysis, decision making, and vehicle control. Its most notable feature is its ability to execute automatic evasive steering maneuvers on crossing pedestrians. It is able to decide within a fraction of a second whether to perform automatic braking or evasive steering, at vehicle speeds typical of urban traffic environment (see  Fig. 29.5).



■ Fig. 29.5

Automatic braking or evasion? That is the question. The system needs to decide within a fraction of a second in response to a suddenly crossing pedestrian

4.2.2 System Description

The Daimler active pedestrian safety system consists of sensor processing, situation analysis, and decision and vehicle-control components. These are now discussed in turn.

Sensor Processing

The sensing component consists of binocular stereo vision (Gehrig and Franke 2007), which has the advantage to provide high-resolution measurements in both the horizontal and vertical direction, as well as an accurate distance map. In order to increase robustness of pedestrian detection, two complementary cues are fused: appearance (pedestrian classification) and motion (moving object detection).

Pedestrian classification utilizes the HOG/linSVM approach of Dalal and Triggs (2005). In order to decide whether a certain rectangular image patch (ROI) represents a pedestrian or not, this approach overlays a spatial grid of cells over the ROI and computes gradient orientation histograms within each cell. A number of local contrast normalization operations are computed, and the resulting normalized histograms are concatenated to an overall feature vector which is used for classification using a linear support vector machine (linSVM). Once an image ROI is confirmed to represent a pedestrian, the distance to the latter is estimated using the computed dense stereo image. Because the exact contour of the pedestrian is unknown within the rectangular ROI, a probability mass function is used for distance estimation, as described in Keller et al. (2010). See ► Fig. 29.6 for some examples of pedestrian classification in urban environment.

Moving object detection involves the reconstruction of the three-dimensional motion field and is performed by the so-called *6D-Vision* algorithm (Franke et al. 2005). This algorithm tracks points with depth known from stereo vision over two or more consecutive frames and fuses the spatial and temporal information by means of Kalman filters. The outcome is an improved accuracy of the estimated 3D positions and of the 3D motions of the considered points. This fusion necessitates knowledge of the motion



■ Fig. 29.6

Pedestrian classification in urban traffic (UK, left driving)



■ Fig. 29.7

Estimation result of the *6D-Vision* algorithm for the moving, partially occluded pedestrian after 0, 80, 120, and 240 ms from first visibility

of the observer, also called the ego-motion. It is estimated from the image points found to be stationary, using a Kalman filter–based approach. Objects are identified as groups of spatially adjacent, coherent motion vectors. Since the *6D-Vision* algorithm not only provides state estimates, but also their uncertainty, the Mahalanobis distance is used as a similarity measure in the cluster analysis. As there is a unique assignment from a tracked 3D point to an object, there is no need to perform an additional object tracking step. However, to increase the robustness, the assignment of the points to the existing object is verified for each frame and new points may be added to the object. See ● Fig. 29.7 for some example output on a partially occluded pedestrian, moving sideways. The *6D-Vision* algorithm already provides after two frames (80 ms) a first estimation result. After three frames (120 ms) there is enough statistical evidence to establish an object hypothesis, even though the pedestrian is mostly occluded by the car in front.

Fusion

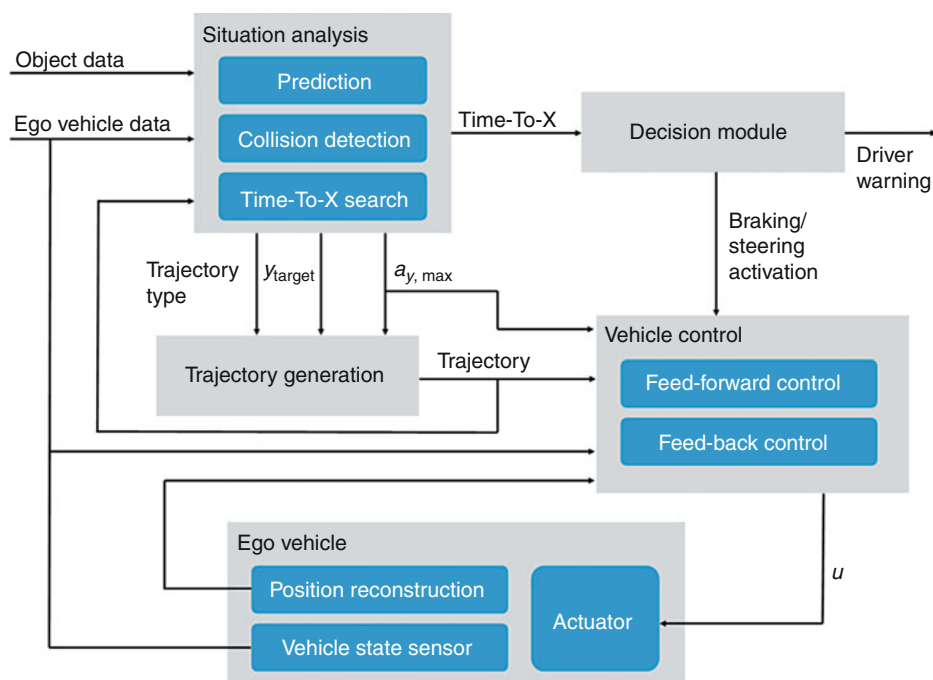
Inputs from the pedestrian classification and *6D-Vision* modules are fused using a Kalman filter. The state S of the filter is given by $S = [x \ y \ v_x \ v_y]^T$ with x/y being the longitudinal/lateral position of the pedestrian to the vehicle and v_x/v_y being its absolute longitudinal/lateral velocity in the world. Constant velocity is assumed for pedestrian motion. Recognitions from the pedestrian detection modules are represented using the measurement

vector: $z_{\text{ped}} = [x \ y]^T$, describing the position of the pedestrian in the vehicle coordinate system. Detections from the *6D-Vision* module contain the velocity and position of possible pedestrian detections. The measurement vector used for these detections is $z_{6D} = [x \ y \ v_x \ v_y]^T$.

As the pedestrian classification module can currently handle only un-occluded pedestrians, fusion with *6D-Vision* is beneficial to initiate tracks quickly in the case of partially occluded, lateral crossing pedestrians, as in [Fig. 29.7](#). As shown in Keller et al. (2010), a further benefit of adding *6D-Vision* to a baseline pedestrian classification system is that lateral velocity estimation is more accurate, which is important for situation analysis.

Trajectory Generation, Situation Analysis, Decision and Intervention, and Vehicle Control

[Fig. 29.8](#) depicts the relationship between trajectory generation, situation analysis, decision and intervention, and vehicle control. Situation analysis predicts how the current driving situation will evolve and automatically evaluates its criticality using measures as, e.g., time-to-collision, time-to-steer, and time-to-brake. This criticality assessment serves as the basis for a decision and intervention module which triggers appropriate maneuvers for collision avoidance and collision mitigation. Such maneuvers are realized by




■ Fig. 29.8

Schematic overview of trajectory generation, situation analysis, decision and intervention, and vehicle control

specialized vehicle controllers. Naturally, vehicle control and situation analysis are closely coupled, since both rely on accurate, realistic models of evasive maneuvers. These models are provided by a trajectory generation module.

Trajectory generation has to provide accurate models of evasive maneuvers that fulfill several requirements: the generated trajectory for evasion should be as comfortable as possible, feasible (i.e., drivable by the ego vehicle), and should also lead to a safe transition with minimal side-slipping of the vehicle during the automatic evasive maneuver. Snatch of steering wheel can be dangerous and therefore must be avoided. Furthermore, trajectory generation should provide the reference input variables for lateral control such as yaw angle, yaw rate, etc. Different trajectory types have been investigated and a sigmoid transfer function based on a polynomial approach was selected to model the evasive maneuver path (Keller et al. 2010).

A numerical simulation method is employed, which allows efficient, real-time computation of Time-To-X criticality measures even for complex maneuvers and which also ensures collision-free evasive maneuvers, if available. As depicted in  Fig. 29.8, the numerical simulation method consists of three main components: prediction, collision detection, and Time-To-X search. In the prediction step, a sequence $\{t_k, z_{ego;k}, z_{obj;k}^1, \dots, z_{obj;k}^M\}$, $k = 1 \dots K$ is computed, where t_k is the k th time stamp of the prediction, K the prediction horizon, $z_{ego;k}$ a vector describing the ego vehicle's pose and motion at time t_k , and $z_{obj;k}^1, \dots, z_{obj;k}^M$ the pose and motion of all M objects provided by the sensor data fusion. These predictions rely on appropriate motion models for all objects and the system vehicle and on assumptions on the ego vehicle's and object vehicles' behaviors. Given the dimensions of all objects in the scene, potential collisions between the system vehicle and objects can be identified by intersecting corresponding positions resulting from $z_{ego;k}$ and $z_{obj;k}^1, \dots, z_{obj;k}^M$ respectively. If a collision is detected, the maximum time step t_k is searched at which a modification of our system vehicle's behavior can still avoid a collision with any of the M observed objects. These time steps are discrete estimates of TTB and TTS and can be found efficiently using a binary search algorithm.

The “decision and intervention” is the core module of the assistance system, since it associates the function with the driver's behavior. Due to the high risk of injuries of a pedestrian in an accident, collision avoidance is the primary objective of the function. In order to identify the best way to support the driver, it is necessary to know the driver's current driving intention. The driver monitoring algorithm uses signals from the vehicle, e.g., accelerator and brake pedal position, speed, lateral and longitudinal acceleration, steering angle, and steering rate, to determine the current driving maneuver of the driver. If the driver is not reacting appropriately to the dangerous situation, an optical and acoustic warning will be given, so he or she can avoid the collision himself or herself. In case a function intervention is necessary to avoid the collision, full braking takes priority over the evasive maneuver. The full braking will be triggered when $TTB = 0$ and the driver is neither doing an accelerating nor an evasive maneuver. Only when the collision cannot be prevented with full braking any more ($TTB < 0$), the evasive maneuver will be activated

at $TTS = 0$ using the vehicle control to compute the necessary steering torque. The function ramps down the steering torque, when the evasive maneuver has finished. Afterward the function is available immediately, when needed. Automatic evasion results in a lateral offset of the vehicle of, e.g., 0.8 m.

Collision avoidance by steering requires precise lateral control of the ego vehicle. The controller permanently compares the reference position along the evasive maneuver trajectory to the actual vehicle position and thus requires accurate and reliable knowledge of the ego vehicle's pose. The position of the vehicle is reconstructed from odometers and inertial sensors readily available in today's vehicles. Using the measured lateral acceleration a_y and the velocity v , the vehicle's heading angle χ can be recovered following

$$\chi(t_k) = \chi(t_k - \Delta T) + \frac{a_y(t_k)}{v(t_k)} \Delta T$$

respectively. Here, ΔT denotes the sampling time step and t_k specifies the time stamp of the k -th iteration step. Using χ and the measured velocity v , numerical integration yields the longitudinal position x and the lateral position y with respect to the current lane

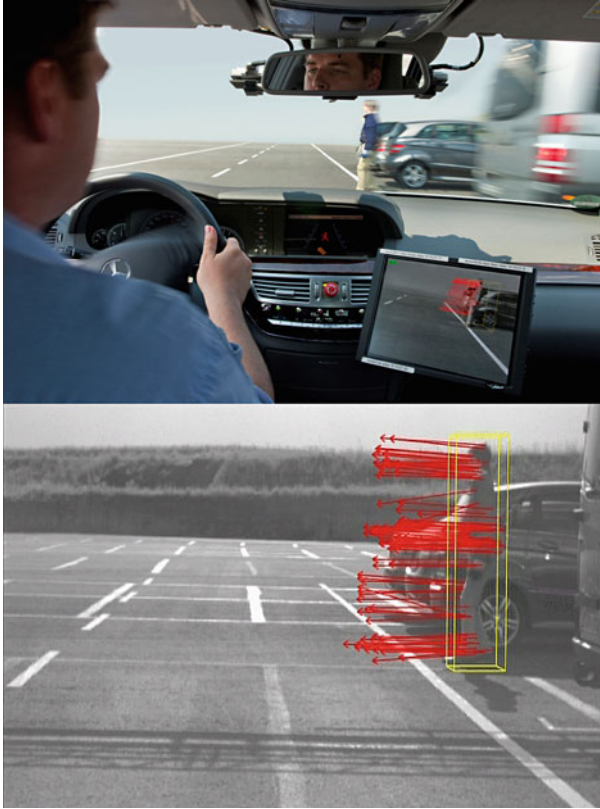
$$\begin{pmatrix} x(t_k) \\ y(t_k) \end{pmatrix} = \begin{pmatrix} x(t_k - \Delta T) \\ y(t_k - \Delta T) \end{pmatrix} + v(t_k) \Delta T \begin{pmatrix} \cos \chi(t_k) \\ \sin \chi(t_k) \end{pmatrix}$$

To account for the nonlinear lateral dynamics of the evasive maneuver, a control strategy combining feed forward and feedback control is used, i.e., the command signal u of the lateral controller comprises the components u_{ff} from a feed forward and u_{fb} from feedback controller, respectively. u_{ff} is computed from the trajectory curvature, that in turn can be derived from the underlying polynomial used. The feedback component u_{fb} is provided by a fourth-order state controller with state vector $(y_{err}; y_{err}^*; \chi_{err}; \chi_{err}^*)$. Here, $y_{err} = y_{trj} - y$ denotes the lateral position error between the reference lateral position and the reconstructed position, $\chi_{err} = \chi_{trj} - \chi$ the difference between reference and reconstructed heading angle. y_{err}^* and χ_{err}^* represent temporal derivatives which can be computed using either derivative lag (DT1) elements, state variable filters, or state observers.

Due to the nonlinear behavior of the vehicle, a gain scheduling approach is employed which adapts both the feed forward gain factor K_{ff} and the feedback gain vector K_{fb} to the current velocity and the maximum allowed lateral acceleration $a_{y,max}$, i.e., $K_{ff} = f(a_{y,max}; v)$ and $K_{fb} = f(a_{y,max}; v)$. For more information, see Fritz (2009).

4.2.3 Experiments

The above mentioned research prototype system was integrated into a Mercedes-Benz S-Class limousine. The vehicle system was tested on a proving ground, where by means of





■ Fig. 29.9

Setup on the proving ground with the pedestrian dummy sliding along a traverse in front of the vehicle. View from inside the vehicle (*top*). Recognized pedestrian including motion estimate (*bottom*)

a traverse construction, a pedestrian dummy, hung by a set of wires, was moved across the road (see ► Fig. 29.9 (top)). An electronic device allowed reproducible movement of the pedestrian dummy. The synchronization of the pedestrian dummy and the vehicle was achieved by a light barrier.

The integrated vehicle system was tested in two scenarios depicted earlier in ► Fig. 29.5. In both scenarios, the vehicle drives 50 km/h and the pedestrian dummy appears from behind an occluding car, with a lateral velocity of 2 m/s. The desired vehicle action is to brake, if still possible to come to a complete standstill, otherwise to evade. It was experimentally determined that the last possible brake time for the vehicle to come to a complete stop corresponds to a pedestrian distance of 20 m (taking into account various device latencies). Similarly, it was experimentally determined the last possible time to safely execute the evasion maneuver to correspond to a pedestrian distance of 12 m. These

distances to the pedestrian could even be shortened, when increasing the total lateral offset from 1 to 2 m and driving the maneuver as depicted in  Fig. 29.2. The resulting braking and steering distances for this maneuver are shown in  Fig. 29.3.

In the first scenario, the pedestrian is first fully visible at about 24 m distance (3.8 m lateral) to the vehicle. This means that the system has only about seven frames (corresponding to 4.1 m driven) to determine pedestrian position and velocity, perform situation analysis, and make the correct decision to initiate braking. In the second scenario, the pedestrian is only first fully visible at about 15 m distance (3.1 m lateral) to the vehicle. This means that the vehicle cannot come to a full stop by braking, therefore the right decision is to evade. For the latter, it has about six frames time (corresponding to 3.5 m driven) to deploy.

Despite the flawless performance on the proving ground, a number of technical challenges remain before this research prototype system can be reliably applied to real traffic. In order to avoid false system activations, the sensing component will need to provide a more accurate pedestrian position and velocity estimation, and deliver increased recognition performance (correct vs. false recognitions). Sensor fusion (e.g., with radar, laser scanners) can provide an important contribution in this regard. The research prototype does not check for oncoming traffic or other obstacles within the commanded driving corridor. Product level systems will additionally require a free space analysis (Badino et al. 2008) to ensure that the automatic evasion maneuver can be safely performed indeed.

5 Conclusion

Steering and evasion assistance systems are a new class of driver assistance systems that open up additional potentials for collision mitigation. It was shown in this chapter that steering intervention is a sensible alternative or additional option for emergency braking systems in a collision speed range above 30 km/h. Steering intervention and evasion systems especially focus on surprising situations, where fast reactions are needed and no time is left for driver warnings. This requires high demands on environment perception as well as on situation analysis. Up to now environment perception algorithms concentrate on object and lane detection and measurement. A new requirement for driver assistance is the detection of free and drivable space, which has to be guaranteed to perform an evasion maneuver.

Three different system layouts were presented: driver-initiated evasion, corrective evasion, and automatic evasion assistance. Driver-initiated evasion only supports an intervention of the driver and therefore offers less safety potential, but due to less complexity it may soon be introduced to market. Corrective or automatic evasion assistance systems are currently investigated by industry and scientific research labs. Beside technical problems like the reliability of the environment perception, a lot of open questions have to be answered, e.g., customer controllability and acceptance. Therefore market introduction is not expected within the next 10 years.

References

- Althoff M, Stursberg O, Buss M (2009) Model-based probabilistic collision detection in autonomous driving. *IEEE Trans Intell Transport Syst* 10: 299–310
- Badino H, Mester R, Vaudrey T, Franke U (2008) Stereo-based free space computation in complex traffic scenarios. In: *IEEE Southwest symposium on image analysis and interpretation*, 2008 (SSIAI 2008), Santa Fe, pp 189–192, 24–26 March 2008
- Barth A, Franke U (2009) Simultaneous estimation of pose and motion at highly dynamic turn maneuvers. In: *DAGM 2009*, Jena, 9–11 September 2009
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, pp 886–893, 20–26 June 2005
- Enzweiler M, Gavrila DM (2009) Monocular pedestrian detection: survey and experiments. *IEEE Trans Pattern Anal Mach Intell* 31(12):2179–2195
- Enzweiler M, Eigenstetter A, Schiele B, Gavrila DM (2010) Multi-cue pedestrian classification with partial occlusion handling. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, 13–18 June 2010
- Fausten M (2010) Accident avoidance by evasive manoeuvres. In: *4. Tagung Sicherheit durch Fahrerassistenz*, TÜV SÜD, Munich, 15–16 April 2010
- Fiorini P, Shiller Z (1996) Time optimal trajectory planning in dynamic environments. In: *IEEE international conference on robotics and automation*, Minneapolis, pp 1553–1558
- Franke U, Rabe C, Badino H, Gehrig S (2005) 6D-Vision: fusion of stereo and motion for robust environment perception. In: *27th DAGM symposium 2005*, Vienna, pp 216–223 (ISBN 3-540-28703-5)
- Fritz H (2009) Verfahren und Vorrichtung zur Kollisionsvermeidung für ein Fahrzeug durch Ausweichen vor einem Hindernis. German Patent Disclosure DE 10 2009 020 648 A1
- Gandhi T, Trivedi MM (2007) Pedestrian protection systems: issues, survey, and challenges. *IEEE Trans Intell Transport Syst* 8(3):413–430
- Gehrig S, Franke U (2007) Improving stereo sub-pixel accuracy for long range stereo. In: *ICCV 07*, Rio de Janeiro
- Hartmann B, Eckert A, Rieth P (2009) Emergency steer assist – Assistenzsystem für Ausweichmanöver in Notsituationen, Internationale VDI-Tagung Reifen-Fahrwerk-Fahrbahn, 12. In: *VDI-Berichte*, vol 2086, pp 131–148
- Hillenbrand J, Spieker A, Kroschel K (2006) A multilevel collision mitigation approach – its situation assessment, decision making, and performance tradeoffs. *IEEE Trans Intell Transport Syst* 7:528–540
- IRTAD (International Traffic Safety Data and Analysis Group) (2006) <http://www.internationaltransportforum.org/home.html>
- Isermann R, Schorn M, Stählin U (2008) Anticollision system PRORETA with automatic braking and steering. *Vehicle Syst Dyn* 46(1):683–694
- Keller C, Dang T, Fritz H, Joos A, Rabe C, Gavrila DM (2010) Active pedestrian safety by automatic braking and evasive steering. *IEEE Trans Intell Transport Syst* (to appear)
- LaValle M (1998) Rapidly-exploring random trees: a new tool for path planning. Technical report, Computer Science Department, Iowa State University
- Neukum A (2010) Controllability of erroneous steering torque interventions: driver reactions and influencing factors. In: *Proceedings of the chassis.tech plus, 1st international Munich chassis symposium*, Munich, pp 365–379, 8–9 June 2010
- Schmidt C, Oechsle F, Branz W (2006) Research on trajectory planning in emergency situations with multiple objects. In: *IEEE intelligent transportation systems conference (ITSC)*, Toronto, pp 988–992
- Schubert R, Schulze K, Wanielik G (2010) Situation assessment for automatic lane-change maneuvers. *IEEE Trans Intell Transport Syst* 11:607–616
- Siegemund J, Pfeiffer D, Franke U (2010) Curb reconstruction using conditional random fields. In: *IEEE intelligent vehicles symposium IV 2010*, San Diego, 21–24 June 2010
- Stiller C, Färber G, Kammel S (2007) Cooperative cognitive automobiles. In: *IEEE intelligent vehicles symposium 2007*, Istanbul, pp 215–220
- Suzumura M, Fukatani K, Asada H (2007) Current state of and prospects for the vehicle dynamics integrated management (VDIM) system. *Toyota Tech Rev* 55(1), March 2007

Velodyne HDL-64E scanner. www.velodyne.com

Werling M, Ziegler J, Kammel S, Thrun S (2010) Optimal trajectory generation for dynamic street scenarios in a frenét frame. In: IEEE conference on robotics and automation (ICRA), Anchorage

Ziegler J, Stiller C (2009) Spatiotemporal state lattices for fast trajectory planning in dynamic on-road driving scenarios. In: IEEE/RSJ international conference on intelligent robots and systems, St. Louis, pp 1879–1884

30 Proactive Pedestrian Protection

Stefan Schramm · Franz Roth · Johann Stoll · Ulrich Widmann
Vehicle Safety Development, AUDI AG, Ingolstadt, Germany

1	<i>Introduction</i>	787
2	<i>Overview of Pedestrian Accidents</i>	788
2.1	International Comparison of Pedestrian Accidents	788
2.2	Analysis of the German In-Depth Accident Study	789
2.2.1	The GiDAS Project: In-Depth Accident Investigation	789
2.2.2	Results of Analyzing Pedestrian Accidents	789
3	<i>Infrastructural and Passive Pedestrian Safety</i>	791
3.1	Infrastructural Pedestrian Safety	792
3.2	Passive Pedestrian Safety	793
3.3	Current Test Procedures for Passive Pedestrian Safety	795
4	<i>Active Pedestrian Safety Systems</i>	796
4.1	Approach of Integrated Vehicle Safety	796
4.2	Overview of Environment Sensor Systems	798
4.3	Functional Algorithm for Collision Prediction	802
4.4	Actuating Elements for Active Pedestrian Safety Systems	805
4.4.1	Brake Assist System	805
4.4.2	Autonomous Braking System	805
4.4.3	Elements for Driver Warnings	806
5	<i>Development of Active Pedestrian Safety Systems</i>	807
5.1	Integrated Development Process of Active Pedestrian Safety Systems	808
5.2	Definition of Active Pedestrian Protection Systems	808
5.2.1	Simulation of Pedestrian Accident Scenarios	809
5.2.2	Informative and Warning System Strategies	810
5.2.3	Autonomous System Strategies	814
5.3	Testing of Active Pedestrian Safety Systems	815
5.3.1	System Test on a Pedestrian Protection Test Facility	815

5.3.2 System Test with the Vehicle-in-the-Loop 817

5.4 Real-World Benefit Calculation of Active Pedestrian Safety Systems 820

5.4.1 Procedures for Calculating the Real-World Benefit 820

5.4.2 Generation of Injury Risk Curves for Pedestrian Collisions 823

6 Conclusion 826

Abstract: Pedestrian accidents are an important aspect of vehicle safety in Europe and throughout the world. Therefore, various countries have already passed statutory regulations on pedestrian protection for vehicles. These mainly focus on assessing passive protection measures. Furthermore the installation of a brake assist system as an active safety system is prescribed in European pedestrian protection legislation. This is because of the significant benefit by reducing the collision speed in pedestrian accidents which was proven in studies with real-world accident data. In future, further active measures will significantly contribute to protect pedestrians because with these technologies it is possible to avoid collisions or mitigate their severity. Combining the active and passive safety technologies to an integrated safety approach will be the most important development objective for further reductions of accidents and casualties in the next years. This chapter provides an overview of active pedestrian protection systems and the new challenges faced when developing these systems. At the beginning, an international comparison of pedestrian accidents and the results of analyzing an in-depth accident database are presented. In the next step, current pedestrian protection measures in the field of infrastructure and passive safety are described. The active safety systems mainly consist of environment sensor systems, functional algorithms, and actuating elements. For each of these components selected, realizations will be shown and discussed for their employment in active pedestrian safety systems. A whole system functionality is created by combining these single modules. In this context, two main system strategies have to be distinguished: on the one hand, system strategies that autonomously engage into the driving situation and on the other hand system strategies that draw the driver's attention to a dangerous situation by presenting a warning. In addition to the development tools new methods for the system test and the benefit calculation have to be engineered. These new tools and test setups will also be presented.

1 Introduction

The protection of pedestrians is an important aspect of vehicle safety in Europe and throughout the world. For example, in Europe (EU-27) approximately 19% of road user fatalities are pedestrians. In the USA, fatally injured pedestrians amount to 11.7% of all fatalities, thus constituting a large proportion of casualties amongst fatally injured road users. Similar figures have been recorded in other countries throughout the world. Therefore, statutory regulations for protecting pedestrians have been passed in many countries. A global technical regulation (GTR) has also been adopted. In addition to the statutory regulations, vehicles' pedestrian protection is also assessed in the New Car Assessment Programmes (NCAP). Within the test specifications of the NCAP and the GTR currently only the effects of passive pedestrian protection systems are taken into account. The European pedestrian legislation also prescribes test procedures for the brake assist system as an active safety system.

The requirements arising from legislation on pedestrian protection and other assessment programs have a considerable impact on the development of vehicles. The vehicle

design and the entire front-end technology are significantly affected. Furthermore, the potential benefits of passive safety measures are limited due to the high mass differences of the parties involved in the collision. Also passive measures do not have an effect on the pedestrians' secondary impacts with objects in the surrounding area like the road. These facts have been taken into account in European legislation for determining the phase 2 pedestrian protection measures. Studies based on real-world accident data have shown that the vehicle equipment with brake assist systems (BAS) leads to a higher reduction in pedestrian casualties than an initially proposed increase of the passive test requirements. This resulted in the mandatory equipment of brake assist systems in Europe since November 2009.

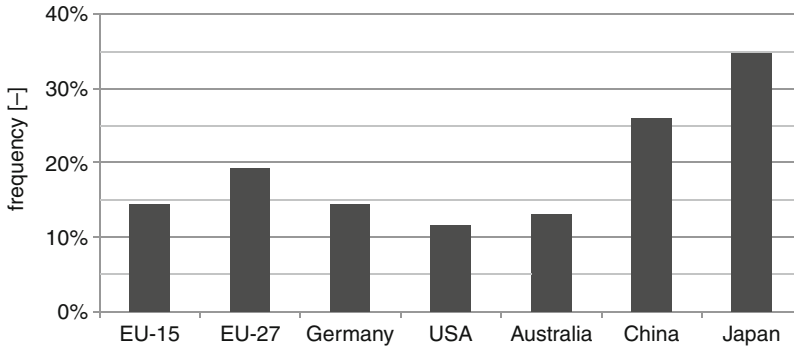
In future further active measures will have a significant share in protecting pedestrians and other road users as these technologies have the possibility to prevent accidents or mitigate their severity. Combining active and passive technologies to an integrated safety approach will be the most important development objective for further reductions of accidents and physical injuries in the next years. This chapter will provide an overview of active pedestrian safety systems and the new challenges faced when developing these systems.

2 Overview of Pedestrian Accidents

For the development of active pedestrian safety systems, it is necessary to analyze pedestrian accidents in order to draw conclusions about how and why these collisions occur and to ensure that the functions are designed on the right focus. Because of that the following section gives an overview about pedestrian accidents on an international level. The accident data collected on a national level are not detailed enough to gain profound information about pedestrian accidents. Thus in-depth accident databases, for example, from the project GiDAS (GiDAS 2010), have to be taken into account to solve this issue.

2.1 International Comparison of Pedestrian Accidents

In the countries of the European Union and throughout the world, pedestrian protection is an important aspect of vehicle safety. There is a significant difference between the number of accidents with pedestrian fatalities in the different countries. In Germany, 14.6% of the total road user fatalities are pedestrians. In the USA, fatally injured pedestrians make up 11.7% and in Japan 34.9% of all fatally injured road users (► Fig. 30.1). Similar high numbers like in Japan can be observed in Korea or Russia. There are several reasons for the significant higher percentages of fatally injured pedestrians, for example, in the Eastern European countries, Japan, or China. This fact is not just the result of vehicle-based influences. Other important factors are also road safety education, infrastructure, high urbanization or rescue.



■ Fig. 30.1

Percentages of pedestrian fatalities of all fatally injured persons in traffic accidents from the year 2008 (SK 2006, CN 2007, EU-27 without LT, CY, BG) (WHO 2009; IRTAD 2010)

2.2 Analysis of the German In-Depth Accident Study

To get more detailed information about traffic accidents involving pedestrians, in-depth accident databases have to be analyzed. One of these databases is built up in the project GiDAS (German In-Depth Accident Study). The following section gives a short explanation of how information about accidents is collected by the GiDAS project. Further an excerpt of the analysis results that can be obtained from this database is illustrated.

2.2.1 The GiDAS Project: In-Depth Accident Investigation

The GiDAS project is a so-called on-the-spot accident investigation research group. It was established in 1999 as a joint venture of the German Automotive Industry Research Association (FAT) and the Federal Road Research Institute (BAST). The accident investigation is carried out within the areas of the two German cities Dresden and Hannover. The GiDAS project is the largest on-the-spot accident study in Germany and collects data from all kinds of accidents with at least one injured person. About 2,000 accidents per year with up to 3,000 accident parameters per collision are investigated following a sample plan which guarantees representativeness to the federal statistics in Germany.

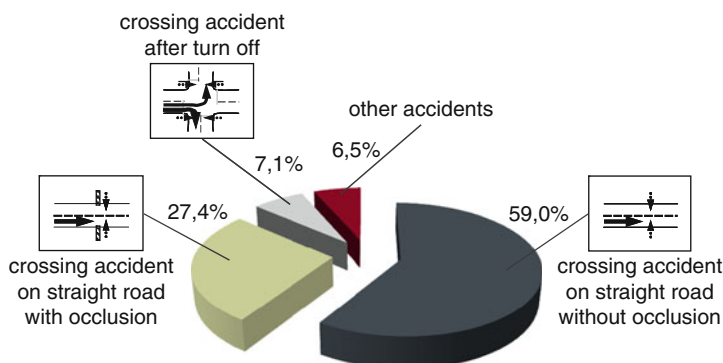
2.2.2 Results of Analyzing Pedestrian Accidents

The pedestrian accidents recorded in the GiDAS accident database contain very detailed information about the reason and the course of the collision events. The findings from an analysis of the accident data are used as input parameters for the development and test of active pedestrian safety systems. Only in this way a high benefit can be expected in real-world accidents. The distribution of accident types in frontal collisions between

passenger cars and MAIS2+ (AAAM 2005) injured pedestrians is shown in [Fig. 30.2](#). The illustration reveals that approximately 86% of all pedestrian accidents occur when a straight road is being crossed by the pedestrian. In 59% of the accidents, the drivers' view was not restricted and the vehicle was driving on a straight road. Only around 7% of the collisions resulted from turning vehicles. The information about the accident types provides a better understanding of how pedestrian accidents are happening. This information is constituted as basis for defining test scenarios on a pedestrian protection test facility ([Sect. 5.3.1](#)) or for the vehicle-in-the-loop test procedure ([Sect. 5.3.2](#)).

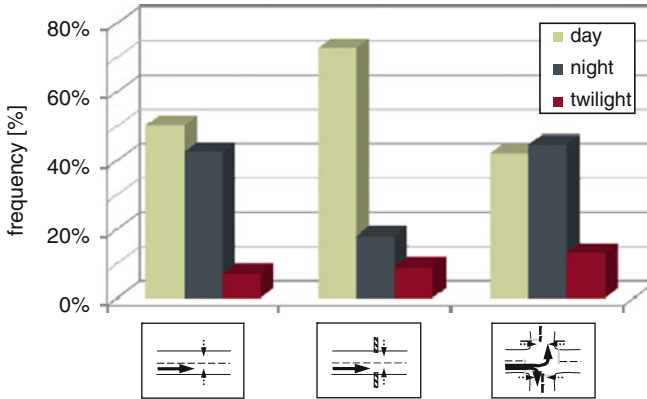
An analysis of the time of day when pedestrian accidents occur shows that 57% happen during the day, 35% at night, and 8% during twilight. The distribution of the time of day within the three accident types from [Fig. 30.2](#) is shown in [Fig. 30.3](#). A comparative analysis of the two accident types on straight roads shows that there is an increase in the proportion of accidents at night for the accident scenario when pedestrians cross the street without an obstruction of view. However accidents on straight roads with restricted view happen more frequently during the day. These findings are also used to define test scenarios for the real-world test of active safety measures. Besides this knowledge also effects the sensor technology that will be chosen to detect pedestrians ([Sect. 5.2](#)).

An analysis of the braking behavior of drivers shows that in approximately 37% of the accidents the collision with the pedestrian occurs without braking by the driver during the precrash phase ([Fig. 30.4](#)). In about 51% of the cases, the driver applied the brakes strongly (deceleration greater or equal to 6 m/s^2). Around 13% of the accidents occur with just a slightly braking maneuver (deceleration less than 6 m/s^2). Especially in collision scenarios with no or little braking of the drivers, it is conceivable that the accident could be influenced positively by, for example, autonomous braking systems or warning the driver. In accidents with braking by the driver, a warning may also induce an earlier reaction. The explanations above make clear that the findings from the real-world accident analysis can



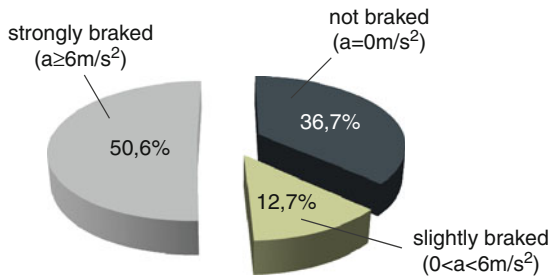
■ Fig. 30.2

Distribution of accident types for frontal collisions between passenger cars and MAIS2+ injured pedestrians (GiDAS accident database)



■ Fig. 30.3

Distribution of accident types for frontal collisions between passenger cars and MAIS2+ injured pedestrians in relation to the time of day (GiDAS accident database)



■ Fig. 30.4

Distribution of the braking behavior of drivers in collisions between passenger cars and MAIS2+ injured pedestrians (GiDAS accident database)

be used both for deriving active pedestrian safety strategies and for defining scenarios for a system test of these safety measures.

3 Infrastructural and Passive Pedestrian Safety

The safety of pedestrians in road traffic is determined by a multitude of different aspects. In the past significant achievements have been made by infrastructural and road safety education measures. The requirements for an aerodynamic vehicle design with smoother vehicle shapes also improved pedestrian protection in the event of an impact. Designing vehicle front ends under consideration of passive pedestrian safety aspects and the development of active safety technologies will significantly improve pedestrian protection in the future.

3.1 Infrastructural Pedestrian Safety

Due to the physical incompatibility of pedestrians and vehicle traffic, it is generally reasonable to separate these two main types of road users. Nevertheless an isolated consideration of pedestrian traffic will not lead to success. In fact all the other road users and their interaction also have to be taken into account. In general, three different classes of infrastructural measures for pedestrian safety are distinguished. These can be applied individually or in combination (FGSV 2002).

The first group is defined as constructional measures. For example, these include a reduction of the lane width in vicinity of pedestrian crossings, the placement of pedestrian traffic refuges, and the reduction of lanes (● Fig. 30.5). These measures lead to a shortened distance that has to be covered by the pedestrians. Further lowering the vehicle travel speed has an important effect on pedestrian safety in road traffic. Studies have shown that road signs alone do not lead to a sufficient reduction of travel speeds. Only constructional modifications of the roadway obtain corresponding results. In this field very effective measures are roadway offsets (● Fig. 30.6) and speed bumps.

Another group of infrastructure measures are operational measures. These include crosswalks and traffic lights. These measures should be applied on roads with a lot of traffic otherwise there is just marginal acceptance by the pedestrians. For roads with a lot of traffic, studies have shown that it is advisable to group pedestrians at intersections and support their intention to cross the street with traffic lights. In addition to constructional and operational measures, other techniques are also feasible. These are, for example, speed monitoring or increasing the driver attention by means of traffic signs on the road or at



■ Fig. 30.5

Long traffic refuge for reduction of vehicle travel speed respectively crossing distance and prevention of overtaking vehicles



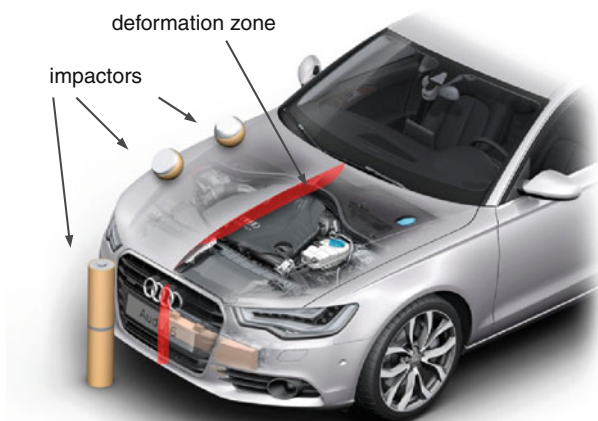
■ Fig. 30.6

Clear characterization of the beginning city limits and an additional roadway offset to reduce vehicle travel speed at a pedestrian refuge

the roadside. Also ensuring that the pedestrians are visible for the vehicle drivers and vice versa improves pedestrian safety by arranging the stationary traffic or preventing sight obstructions at intersections (Maier 1984).

3.2 Passive Pedestrian Safety

In general, pedestrian collisions constitute of two phases. During the primary impact, the pedestrian crashes with the vehicle. This primary impact is followed by a secondary impact with objects around the vehicle, especially with the road surface. The vehicles' passive safety measures help protecting the pedestrian during the primary impact. For a minimization of injury severity in the event of an impact at the vehicle, large-area force application is beneficial in order to convert kinetic into deformation energy. So all components in the front end are developed with respect to stiffness and deformation capability. The distance between bonnet and the package components in the engine bay is just one example (► Fig. 30.7). The design of the bonnet, its add-on parts such as hinges, locks, gas springs, and the windscreen wipers with their electric drives are also affected as well as fenders and the surrounding structural parts. Another objective of the passive measures is the prevention of large relative movements within certain parts of the body, for example, the lower leg or the knee. For this reason, the bumper is integrated into the vehicle front in order to limit the bending and transverse loads in the event of a leg impact (► Fig. 30.7). Furthermore, deformation elements and zones are implemented into the front end to minimize the applied load on the human leg.



■ Fig. 30.7

Deformation zones between bonnet and engine bay package and at the front bumper with pedestrian headform and lower leg impactors

One can clearly see that the requirements of the passive pedestrian protection measures have significant influence on vehicle development. The design, the technology, and the concept of the front end are strongly affected. These circumstances often lead to conflicts of objectives and to limitations of other component functionalities. For example, at the bonnet the necessarily flexibility for pedestrian protection conflicts with the required stiffness for high wind loads. Implementing foam into the front end affects the available space for the crash boxes which are located in front of the frame side members. The deformation zones required above the fender bench reduces its cross section and makes it more difficult to connect it with the A-pillar. Increasing the deformation zones in the area of the bonnet can also result in negative effects on aerodynamics due to the increased vehicle dimensions.

In the future, more and more crash-active measures will be found in the front end of vehicles. These systems are, for example, pop-up bonnets which enlarge the deformation area when a pedestrian collision occurs. Pedestrian airbags that address the hard structures, for example, in the area around the A-pillar, are also possible. Adding foam and crash-active safety measures on the vehicle results in a higher weight. Although the passive measures have a positive effect on protecting pedestrians in the event of a collision, their safety potential is limited. This is also shown in studies by Liers (2009) and Hannawald and Kauer (2003) which were performed on the basis of real-world accident data. One reason for this is the large difference in mass of the two collision partners. Secondly the application of energy depending on the collision speed as main factor for the pedestrians' injury severity is not addressed by the passive safety measures. Furthermore, there is no possibility for the passive measures to positively affect the pedestrians' secondary impacts.

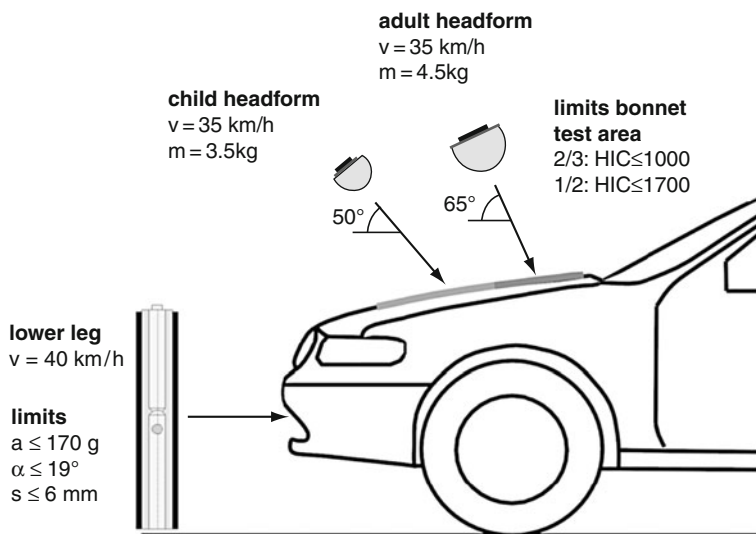
3.3 Current Test Procedures for Passive Pedestrian Safety

Before new vehicle models are permitted to be sold on the market, the manufacturers have to prove that they comply with the statutory requirements on vehicle safety. Since 2005, statutory test regulations for pedestrian protection exist in Japan and Europe. These regulations must be complied with for a homologation of new vehicle models. In addition to the statutory requirements, there are also consumer tests such as the NCAP.

These also assess the safety level of new vehicle models. Contrary to the statutory regulations, the consumer tests aim to conduct a comparative assessment of different vehicle models in order to help the customer choose particularly safe vehicles. The test criteria and limits are often considerably higher compared to the statutory requirements and also included in the vehicle manufacturers' technical specifications. In addition to the traditional vehicle crash tests for the assessment of passenger protection, the test regulations of the NCAP in Europe (ENCAP), Japan (JNCAP), and Australia (ANCAP) also include test procedures for an evaluation of passive pedestrian safety on the vehicle. Furthermore a global technical regulation was passed at the end of 2008 (ECE 2009).

Analyses have shown that injuries due to collisions between passenger cars and pedestrians mainly occur at the legs and head. Therefore, the test procedures and load limits both in legislation and consumer test organizations have been developed to assess safety measures on the vehicle protecting the lower extremities, the pelvis, and the head. The statutory test regulations and those in the NCAP are aimed at assessing the passive pedestrian protection measures (➤ Sect. 3.2). Component tests are carried out with impactors (➤ Fig. 30.7) to evaluate the quality of these measures. The impactors are models of the corresponding body parts listed above. In order to determine whether defined load limits are complied with in the event of an impact with the vehicle's front end, the forms are fired at the vehicle at various speeds and impact angles and the individual loads are measured.

At the beginning of 2009 the directive EC78/2009 was passed for Europe (European Parliament 2009). It describes the enhanced statutory pedestrian protection requirements phase 2 that must be complied with from 2013 onward. ➤ Figure 30.8 illustrates the European phase 2 component test for pedestrian protection. A study carried out on behalf of the European Commission in the context with defining the new pedestrian legislation phase 2 has shown that pedestrian protection can be improved significantly by a combination of passive and active measures. In particular, the study indicates that brake assist systems have better pedestrian protection potential than tightening the passive requirements from the first statutory phase (European Commission 2007; European Parliament 2009; Hannawald and Kauer 2003). Therefore, the European phase 2 directive also specifies that from November 2009 onward all vehicles must be equipped with a brake assist system in order to gain EU type approval. The directive EC78/2009 is the first legislation in the world that prescribes the implementation of both passive and active vehicle safety measures. Further the directive also announces that vehicles equipped with a collision avoidance system could be exempt from certain requirements of the directive. This derives from the fact that collision avoidance systems have the potential to prevent



■ Fig. 30.8

Component test of the European legislation phase 2 with load limits for the impactor forces

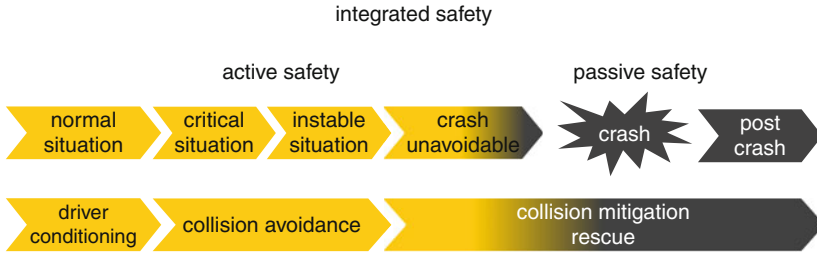
collisions with pedestrians rather than just lower the consequences in pedestrian injury severity. Article 11 of the directive (European Parliament 2009) states that if equal effectiveness of collision avoidance systems is proven it is not necessary to comply with the requirements on passive protection measures. A test procedure for this assessment has not been defined yet.

4 Active Pedestrian Safety Systems

Active safety systems differ from passive ones in that their effects are deployed before the actual collision situation. On the one hand active measures may intervene in critical situations and possibly prevent a potential collision. On the other hand these active systems may mitigate the severity of the accident when the collision is unavoidable. In general, active safety measures consist of environment sensor systems, functional algorithms, and actuating elements. In conjunction with passive pedestrian safety features, these active measures constitute the integrated safety approach.

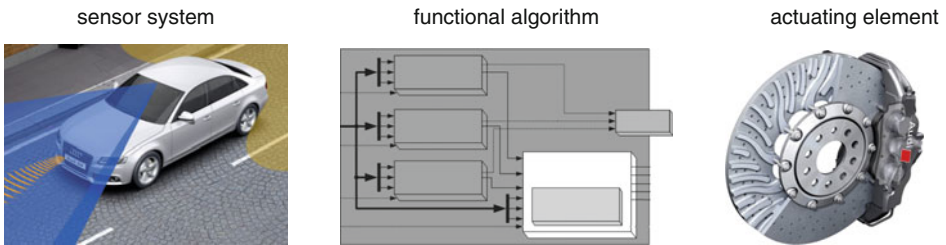
4.1 Approach of Integrated Vehicle Safety

Vehicle safety is moving toward an integrated view in which the collision situation is considered in its entirety. The focus is no longer just on the actual collision but also on the phases before and after it. ► Figure 30.9 illustrates that the integrated safety combines



■ Fig. 30.9

Combination of active and passive safety measures to an integrated approach of vehicle safety



■ Fig. 30.10

Schematic components of active safety systems

active and passive safety systems. The strategy of the active components is to prevent the occurrence of a critical situation by means of driver conditioning by assistance and information systems during the normal driving situation.

If the normal driving situation is left, it may still be possible to avoid a collision by warning the driver or autonomous interventions in the vehicle dynamics. Should the situation develop in such a way that the accident is no longer avoidable it is possible to reduce the severity of the accident, for example, via autonomous braking systems. Furthermore in this phase adaptive passive protective systems can also be prepared for the upcoming collision in order to provide additional protection. If a collision occurs, the effect of passive safety systems reduces the consequences of the accident. Calling the emergency services automatically after a collision is also allocated to the integrated safety approach. This general description of integrated safety can also be found within the field of pedestrian protection. Therefore, the components of active pedestrian protection measures will be described in more detail in the following sections. The passive safety measures have already been described in ▶ Sect. 3.2.

In general, active safety measures consist of environment sensor systems, functional algorithms, and actuating elements. These elements are shown in ▶ Fig. 30.10 schematically. The signals detected by the sensor systems are forwarded to a functional algorithm.

On the basis of defined input signals, this algorithm calculates whether the vehicle is in a situation in which actuating elements should be activated.

4.2 Overview of Environment Sensor Systems

The decision for triggering active safety elements on the vehicle is made by functional algorithms which receive defined sensor signals as input parameters. This includes information on the status of the vehicle as well as on the environment. The section below will focus on sensor systems for detecting pedestrians in the environment of the vehicle. ➤ [Figure 30.11](#) contains a list of selected sensor systems currently available for use

sensor system	characteristics
mono camera	<ul style="list-style-type: none"> • no direct distance information • good object classification possibilities • restricted performance at night
stereo camera	<ul style="list-style-type: none"> • distance information • good object classification possibilities • restricted performance at night
far-infrared camera	<ul style="list-style-type: none"> • good night performance • good object classification possibilities • restricted performance in daytime
radar sensor	<ul style="list-style-type: none"> • long range • good detection of distance and relative velocity in longitudinal direction • less speed and position resolution in lateral direction • relatively small radar cross-section of the pedestrian
laser scanner	<ul style="list-style-type: none"> • long range • large angle of aperture • independent from ambient light
photonic mixer device (PMD)	<ul style="list-style-type: none"> • direct distance measurement for each pixel • fast 3D recording with high frame rate • active lighting system • independent from ambient light

■ Fig. 30.11

Examples and characterizations of available environment sensor systems

in vehicles which will be described and discussed for their usage in pedestrian protection. The application of camera systems is one option for detecting pedestrians around the vehicle. For example, it is possible to identify a pedestrian in traffic situations using camera images provided by a mono camera. The mono camera projects a three-dimensional scene onto a two-dimensional image. As a consequence, it is hardly possible to draw conclusions about the three-dimensional geometry of the traffic situation on the basis of the two-dimensional image. This means that no distance values are directly available for the individual image elements or classified objects. To determine the distances for objects in a two-dimensional image, this sensor system is usually combined with other sensor systems. For example, it may be feasible to combine the mono camera system with a radar sensor or laser scanner (● Sect. 5.2.3). If there is no combination with other sensor technologies, it may also be possible to calculate the depth information out of image sequences.

This challenge can be solved by a stereo camera system. This system consists of two cameras fitted side by side which provide two images of the environment with a different viewing angle. Information on the depth of image elements and objects can be calculated from these two camera images. The main advantage of camera systems is that they make it possible to identify objects by their appearance or their kind of movement. As the performance is influenced to a large extent by the ambient light, these systems have shortcomings when used in twilight and at night. Infrared cameras are particularly well-suited for use during the night. A far-infrared camera is shown in ● Fig. 30.12. It picks up thermal radiation from objects within its field of vision and displays the differences in temperature in a black and white image. The pictures of a scene with headlights at night and in the display of the night vision assistant of the Audi A8 which is based on the far-infrared camera system in ● Fig. 30.12 is shown in ● Fig. 30.13.

The use of a radar sensor in conjunction with a mono camera has been mentioned above. Radar sensors emit electromagnetic waves that are reflected from almost all



■ Fig. 30.12

Far-infrared camera of the night vision assistant in the Audi A8 (Taner and Rosenow 2010)



■ Fig. 30.13

Scene with headlights (*left*) and in the display of the night vision assistant of the Audi A8 based on a far-infrared camera system (*right*)

surfaces but especially from metallic objects. The reflected waves are picked up again by a receiver. The distance to the objects can be measured from the duration of these electromagnetic waves. Relative speeds between the transmitter and the detected objects are usually measured via the Doppler effect. The main features of radar sensors are the long range and good speed and position detection in longitudinal direction. These sensors are only suitable for use in active pedestrian protection systems to a limited extent unless they are combined with other sensor systems, for example, a mono camera. This is due to the relatively small radar cross section of the pedestrian and the speed and position resolution in a lateral direction to the vehicle.

The laser scanner is a system whereby a laser emits light pulses and the light is picked up again. Depending on the number of beam layers under circumstances only limited height information can be provided by this type of sensor system. Laser scanners usually have large angles of aperture and long ranges. As with radar sensors, these systems are independent upon the ambient light.

A sensor system that provides image and distance information with just one single picture is the PMD (photonic mixer device) sensor. As opposed to the sensor systems mentioned before, this kind of three-dimensional detection provides information on the absolute geometrical dimensions of objects independent from the kind of surface or ambient light. A PMD sensor system consists of an active lighting system and a receiver as shown ● Fig. 30.14.

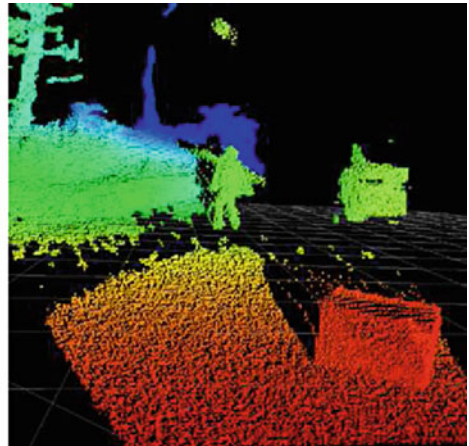
The sensor parameters such as the angle of aperture, range, and frame rate can be applied using these two components. The significant advantage of PMD sensor systems is that distance information is available for each recorded pixel. The PMD receiver is able to take pictures at a frame rate far in excess of 50 images per second.

That ensures that the type of sensor complies not only with the requirements on reliability and good packaging but also the standards required for detecting highly



■ Fig. 30.14

PMD sensor system from ifm electronic gmbh consisting of an active lighting and a camera system





■ Fig. 30.15


Two similar scenes recorded in 2D (left) and with a 200×200 pixels PMD camera system (right) (PMDTec 2010)

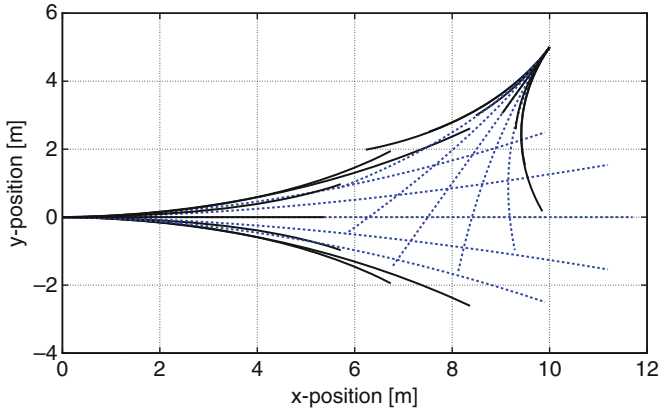
dynamic movements. A pedestrian scene and the image of a PMD sensor system are shown in ► Fig. 30.15. The image of the scene is taken with a camera resolution of 40,000 pixels. This is a future upgrade of the PMD system. Currently PMD sensors for automotive use are available with a resolution of 1,024 pixels.

4.3 Functional Algorithm for Collision Prediction

Functional algorithms are used to calculate whether a dangerous situation or a potential collision will occur depending on the movements of the vehicle and the pedestrian. If the vehicle and the pedestrian are on collision course at first a risk threshold is reached as the two objects approach each other. In the next time steps, there will be a point when the collision can no longer be avoided. None of the actions taken by the driver and the pedestrian from this point onward will prevent the collision from happening. One possible approach for determining the point in time when the accident is unavoidable is presented in (Botsch and Lauer 2010). In the subsequent section, this collision mitigation algorithm is described. It is important to point out that in this paper the algorithm is explained for unavoidability determination involving two vehicles. But the general approach can also be used for vehicle-pedestrian accidents.

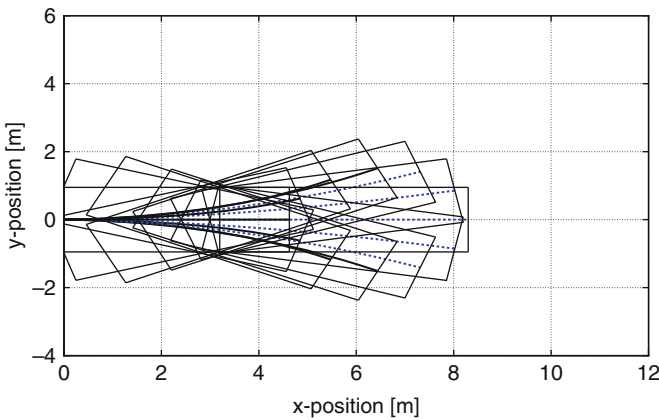
In the algorithm presented by Botsch and Lauer (2010) all possible future trajectories for the ego vehicle and for the object in the vicinity of the ego vehicle are computed. If all these trajectories lead to a collision, it can be concluded that a crash is unavoidable. The number of trajectories that must be computed can be reduced by considering only extreme trajectories resulting from maximum accelerations or decelerations in tangential or normal direction. These accelerations or decelerations can be approximated using Kamm's circle. This model is based on the fact that there is a maximum friction force between the vehicle's wheels and the road. To predict the trajectories of objects, a set of differential equations has to be solved. As initial conditions for the differential equations, the current values of the vehicle's location, velocity, and heading angle are taken into account. Whereas the values for the initial conditions required to calculate the extreme trajectories of the ego vehicle are available for all other objects, adequate attributes have to be determined by coordinate transformations. For this step, the ego velocity and the ego yaw rate are taken into account. After these transformations, the set of differential equations can also be solved for the objects in the vicinity of the ego vehicle, and the extreme trajectories of the objects around the ego vehicle can be predicted within the same coordinate system in which the motion of the ego vehicle is predicted. As an example, the predicted trajectories of the ego vehicle and another vehicle for the next 800 ms are illustrated in  Fig. 30.16. The number of computed trajectories in  Fig. 30.16 is reduced to 12 both for the ego vehicle and the other object. The dashed curves represent trajectories caused by acceleration and the solid lines represent trajectories caused by deceleration.

In order to check whether a crash will occur or not, rectangular boxes are placed on the predicted trajectories at every future time stamp. For a time stamp which is 600 ms in the future for the ego vehicle, this representation is shown in  Fig. 30.17. Given one ego vehicle trajectory and one object trajectory, a graph such as this can verify if at least one corner of the object lies within the area described by the ego vehicle for each time instance and if a collision will occur. In the collision decision algorithm, a check is carried out for all possible combinations of ego vehicle and object trajectories whether a collision will occur in the near future. A maximum prediction time (e.g., 2 s) must be set. The collision



■ Fig. 30.16

Predicted trajectories of two vehicles assuming a friction of 10 m/s^2 (Botsch and Lauer 2010)



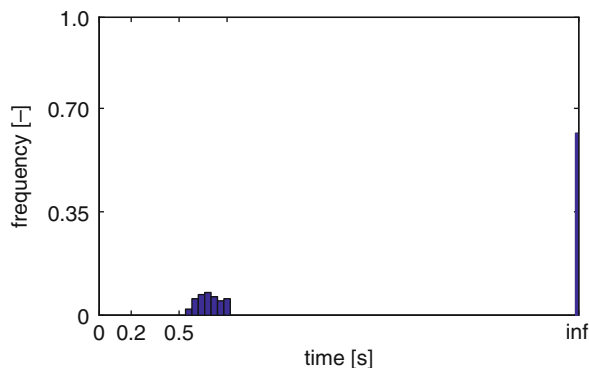
■ Fig. 30.17

Possible positions of the ego vehicle in 600 ms (Botsch and Lauer 2010)

is considered as unavoidable if all possible combinations of ego vehicle and object trajectories lead to collisions.

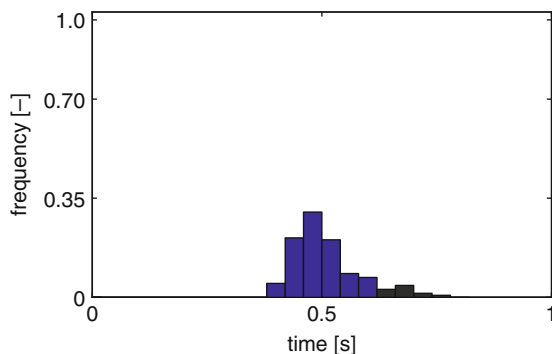
This kind of algorithm not only decides whether a collision is unavoidable but also estimates the conditional probability density function of the time to collision (TTC) random variable assuming maximum acceleration for the ego vehicle and the objects. The conditional probability density function of the TTC gives important insight into the criticality of a scenario since the time interval in which a collision might occur is known.

In ► Fig. 30.18 the estimated conditional probability density function of the TTC for the scenario in ► Fig. 30.16 is shown. As it can be seen there are some trajectories which lead to collisions having a mean TTC of approximately 700 ms. However about 60% of all 144 considered trajectories do not lead to crashes (grouped as “inf” in ► Fig. 30.19) and



■ Fig. 30.18

Estimated TTC probability density function for “collision avoidable” (Botsch and Lauer 2010)



■ Fig. 30.19

Estimated TTC probability density function for “collision unavoidable” (Botsch and Lauer 2010)

therefore the scenario is not classified as unavoidable. The result of another scenario is shown in [Fig. 30.19](#). In this case, the situation is classified as unavoidable since the initial conditions of the two vehicles lead to a TTC probability density function in accordance with [Fig. 30.19](#).

The collision mitigation algorithm described above assumes worst-case conditions by imposing no constraints on possible trajectories on the driver’s speed of reaction, on the weather conditions, etc. For example, if it is known that there is a guard rail on the right of the ego vehicle some of the possible trajectories can be excluded. If such knowledge is available, the algorithm can be easily adapted leading to an estimated conditional probability density function of the TTC, which reflects reality more accurately. Based on the principles above, a similar algorithm can be easily derived to estimate a collision probability and thereby to anticipate different risk levels.

4.4 Actuating Elements for Active Pedestrian Safety Systems

From the data provided by the environment sensors and the vehicle internal sensor systems, the functional algorithm calculates whether defined actuating elements should be activated. A differentiation must be made between actuating elements for longitudinal dynamics (e.g., braking) and lateral dynamics (e.g., steering). Warnings to the driver are also issued via defined warning actuating elements. The next few sections provide an overview of selected actuating elements for use in active pedestrian safety systems.

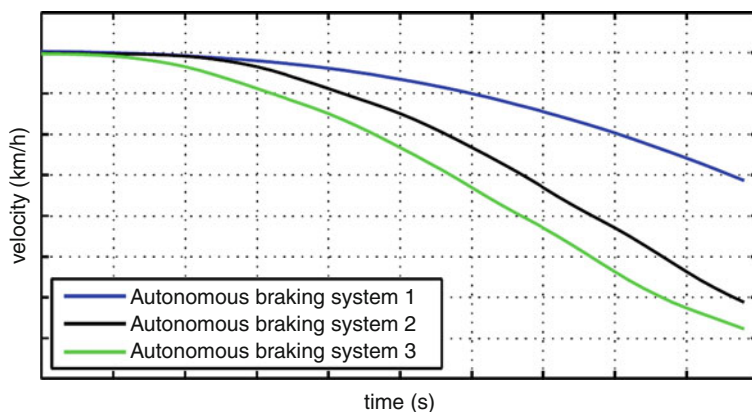
4.4.1 Brake Assist System

Brake assist systems (BAS) support the driver when making emergency stops. The brake assist system detects that the driver wants to make an emergency stop and it automatically increases the vehicle's deceleration to the maximum level that is physically possible. Normal drivers are therefore able to decelerate and brake in a way that only very experienced drivers would otherwise be capable of. The brake assist system is thus an actuating element that is not activated autonomously but is triggered by the driver. An emergency stop is activated on the basis of the way in which the driver actuates the brake pedal. Other parameters such as the vehicle speed are often also taken into account for activation. The activation of the brake assist system facilitates optimal deceleration, braking distance, and speed reduction. There are various ways for realizations of brake assist systems on the vehicle. For example mechanic, electronic, and hydraulic brake assist systems are available, which are described more detailed in Heiing and Ersoy (2008). The significant benefits provided by a brake assist system in real-world pedestrian accidents are presented in Hannawald and Kauer (2003) and Schramm (2011) for example. In consequence of the high level of effectiveness of this system with respect to unprotected road users in general and pedestrians in particular, the installation of such systems has become mandatory in European legislation since the year 2009.

4.4.2 Autonomous Braking System

Modern cars are already equipped with driver assistance systems and safety functions that automatically affect the longitudinal dynamics of the vehicle. Opposed to brake assist systems, these autonomous braking systems are deployed to decelerate the vehicle without driver intervention. In this context, the possibility of oversteering by the driver has to be taken into account. The braking pressure of these autonomous systems is usually built up by a hydraulic unit. This is the key component because the hydraulic unit converts and implements the control commands from the control unit. This is usually performed by electromagnetic valves which control the pressure on the wheel brakes.

Activation of an autonomous braking system causes the vehicle to decelerate. This deceleration results in a defined reduction in velocity. ► [Figure 30.20](#) shows three



■ Fig. 30.20

Schematic velocity-time graphs of autonomous braking systems with different performances in reducing the initial vehicle velocity

velocity-time curves. One can see that different reductions in velocity are achieved by different autonomous braking systems. These differences are caused, for example, by the basic technical design of the braking systems or whether the system is prefilled before the braking operation is triggered. As explained in ▶ Sects. 5.2.2 and ▶ 5.2.3 there are high requirements on the braking systems' reaction and pressure buildup times in order to achieve a significant reduction in velocity especially if the system is triggered in a short time before collision.

4.4.3 Elements for Driver Warnings

In addition to the use of braking systems, also driver warnings can be implemented as an actuating element for active pedestrian safety measures. These warnings can be conveyed to the driver via the visual, acoustic, kinesthetic, or tactile sensory channel. An assessment of the quality of these four sensory channels with respect to the amount of information that can be communicated and the time needed for a reaction is shown in ▶ Fig. 30.21.

There is a wide variety of possible concepts for the design of warning actuating elements. For example, the driver can be notified of detected objects that are on potential collision course on the display of the instrument cluster (▶ Fig. 30.22) or by means of an acoustic warning signal.

Another method of drawing the driver's attention to an upcoming collision with a pedestrian is to project a visual warning onto the windscreen which represents the brake lights of a vehicle ahead (▶ Fig. 30.23). Triggering a brake jerk is one example for a kinesthetic warning. Tightening the seat belt and steering wheel vibration are warnings that address the tactile sensory channel. In general, warning elements have to be investigated more closely for use in a vehicle in order to provide the driver the best possible

sensory channel	amount of information	reaction time
visual	very high	fast
acoustic	medium	medium
kinaesthetic	low	very fast
tactile	low	very fast

■ Fig. 30.21

Human sensory channels with the amount of information that can be communicated and the time taken to react to a signal (Hoffmann and Gayko 2009)



■ Fig. 30.22

Visual pedestrian warning in the display of the night vision assistant of the Audi A8 (Taner and Rosenow 2010)

assistance in each particular situation. In this context, it is referred to [Sect. 5.2.2](#) for more information about informing or warning system strategies.

The information provided above clearly shows that a multitude of technical solutions are available for driver warning actuating elements, which can also be used in combination. Generally speaking, warnings are intended to assist the driver in selecting a suitable response to avoid a potential collision situation. It is therefore essential that studies are carried out to examine the effectiveness of warnings in dangerous or collision situations.

5 Development of Active Pedestrian Safety Systems

The development of active pedestrian protection systems requires new processes, tools and test procedures. The following section will present an approach to define and test active pedestrian protection technologies. An important part of this development process also is the real-world benefit evaluation.



■ Fig. 30.23

Visual pedestrian warning of the manufacturer Volvo by projecting a red light bar in to the windscreen (Volvo 2010b)

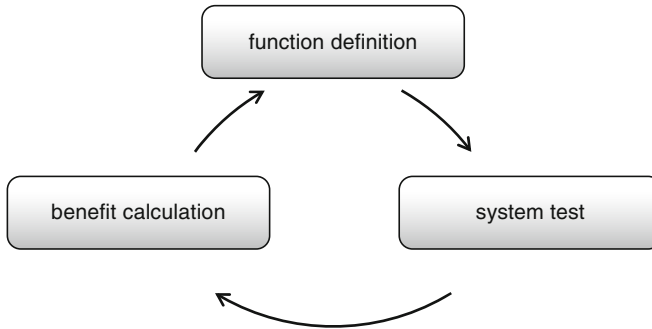
5.1 Integrated Development Process of Active Pedestrian Safety Systems

The development of active safety functions focused on pedestrian protection requires an integrated development process. This process consists of function definition, system test, and benefit calculation (► Fig. 30.24). In the step of function definition, the concepts of actions and the system components necessary are determined (► Sect. 5.2).

The system test involves analyzing the single system components or the entire functionality both in simulation and in close to reality scenarios, for example, on a pedestrian test facility (► Sect. 5.3.1). The process step of benefit evaluation involves an assessment of the safety level provided by the entire system with respect to the potential of collision avoidance and mitigation in real-world pedestrian accidents taking into account the findings of the two process steps mentioned before. The three stages function definition, system test, and benefit calculation are closely linked with each other and build up an iterative process in the development of active pedestrian safety functions. Only in this way, it is possible to design a system functionality with a high real-world benefit.

5.2 Definition of Active Pedestrian Protection Systems

The function definition is one phase of the integrated development process (► Fig. 30.24). During this phase, the sensor systems, functional algorithms, and actuating elements are built to a total system. In general, there are two system approaches. On the one hand, it is possible to intervene in the vehicle dynamics autonomously without including the driver.



■ Fig. 30.24

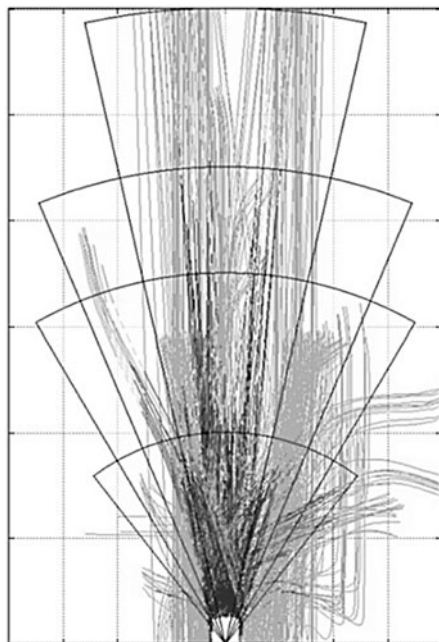
Integrated function development process consisting of function definition, system test, and benefit calculation

An example of this is triggering an emergency brake if a collision is unavoidable. On the other hand, it is possible to draw the driver's attention to a critical situation by means of a warning element. Also combinations of the two main approaches are imaginable.

5.2.1 Simulation of Pedestrian Accident Scenarios

The definition of active pedestrian protection system strategies is based on the findings from the accident analysis, as shown in 🔍 Sect. 2.2.2. Based on the statistical distribution of the pedestrian accident types and the collision parameters from the GiDAS database, the collision situations are rebuilt in a simulation environment. By stochastic distributions of the accident parameters, it is possible to create an extensive simulation database that contains collisions as well as non-collisions. In this way, it is possible to statistically construct traffic scenarios or accident events by selecting appropriate accident simulation scenarios as these are observed in reality. For the function definition and the system test, it is essential that both collisions and non-collisions are available within the simulation database because information on a system's false positive rate cannot be collected on the basis of just collision situations. This simulation database makes it possible to carry out basic investigations for the use of different system components in the subsequent steps. However, only those scenarios that lead to a collision between the vehicle and the pedestrian are used for the benefit calculation in 🔍 Sect. 5.4. Further these simulation scenarios are copies of the real-world accidents such as those recorded in the GiDAS database.

For the determination of requirements for the sensor systems, the sensors are modeled in the simulation environment. After that the detection rate of different sensor systems can be analyzed. These studies provide information about the point in time when the pedestrians appear in the sensors' field of view or are detected by the systems depending on different parameters such as, for example, the angle of aperture, range, and update rate (🔍 Fig. 30.25). The simulation results show that generally large angles of aperture are



■ Fig. 30.25

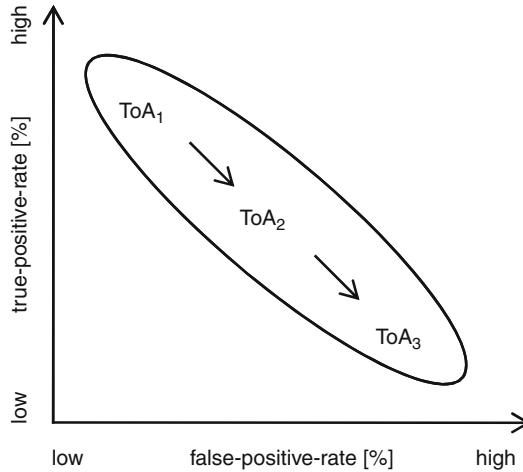
Analysis of various sensor fields of view and sensor ranges in the simulation for different pedestrian trajectories

needed to achieve high detection performance which is derived from the huge amount of crossing accidents (► Fig. 30.2). In this context, one must bear in mind that a separate analysis of the sensor system without reference to the functional algorithm only provides limited information because the intended trigger times of the actuating elements also have an effect on the sensor parameters. Therefore, prototypes of the functional algorithms are also implemented in the simulation environment. In general, the following tendency was found out: the earlier an actuating element should be triggered before collision the higher is the requirement at the sensor range.

A further result of a combined analysis of sensors and prototype functional algorithms is also shown in ► Fig. 30.26. This illustrates that the further the time of action is brought forward from the point of collision, the higher is the false positive rate. This result derives from the fact that pedestrians are highly dynamic objects in comparison to vehicles and therefore have the opportunity to enter or leave the dangerous zone before the actual vehicle impact up to a short time before collision.

5.2.2 Informative and Warning System Strategies

System strategies whereby the driver is included to positively influence an upcoming collision situation involve actuating elements that warn the driver (► Sect. 4.4.3). Because



■ Fig. 30.26

Relationship between the true positive and false positive rate depending on the time of the action (ToA) with $ToA_1 < ToA_2 < ToA_3$

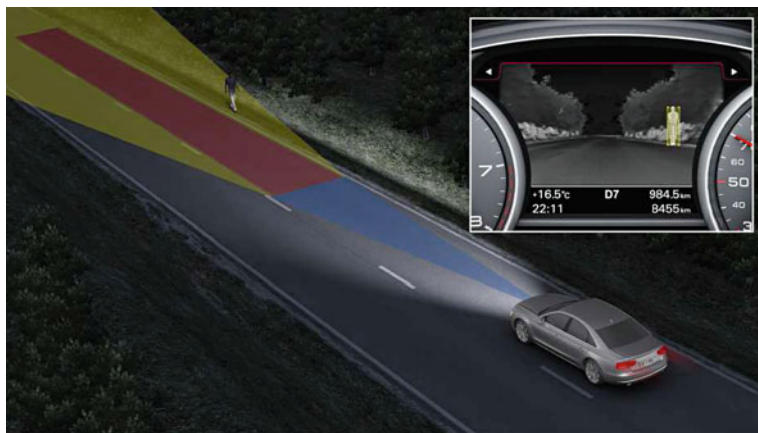
of the driver's reaction time these concepts require sensors with relatively long ranges in order to be able to warn the drivers in sufficient time before the collision. In comparison to an automatic reaction to an input signal, humans have a relatively long reaction time so that warning must be given a relatively long time ahead of the potential collision situation. In this context, the connection between the detection and false positive rate shown in [Fig. 30.26](#) also has to be considered.

When developing a warning system, some general principles of design must be kept in mind. On the one hand the connection between the warning and the reason for it has to be understood by the driver. Due to the increasing number of assistant systems providing information for the driver, it is also necessary that the system issuing the warning can be identified quickly and accurately. Furthermore it must be possible to identify the warning clearly under all conditions (e.g., different light conditions or background noises). Finally, it is essential that the warning draws enough attention to itself without obscuring the reason for it.

A warning can be given to the driver using various sensory channels, for example, via visual, acoustic, kinesthetic, or tactile input ([Sect. 4.4.3](#)). The realization of a warning system in the vehicle not only requires theoretical investigation based on the criteria mentioned before but also real-world tests with probands. The vehicle-in-the-loop test procedure explained in [Sect. 5.3.2](#) can be used to carry out such real-world trials.

Pedestrian protection warning systems are already installed in modern vehicles. The night vision assistant is one example. It uses a far-infrared camera ([Fig. 30.12](#)) that generates an image of the scene in front of the vehicle. In this image, warm objects are light-colored and cold objects are dark. Due to the emitted body heat of the pedestrians, these objects can be seen particularly well against the mostly low-temperature background

allowing detected pedestrians to be displayed at a distance of up to 100 m. The far-infrared camera also provides an image of the road and outlines of buildings. If the functional algorithm identifies that there is an upcoming collision between the vehicle and a detected pedestrian, the person which was initially marked yellow (► Fig. 30.27) is then illustrated in a red box (► Fig. 30.28) and an acoustic warning signal is issued from the instrument cluster.



■ Fig. 30.27

Detected pedestrian not on potential collision course with the vehicle marked in highlighted color in the instrument cluster display (Taner and Rosenow 2010)



■ Fig. 30.28

Detected pedestrian on potential collision course with the vehicle marked in highlighted color in the instrument cluster display (Taner and Rosenow 2010)

Another active pedestrian protection strategy is the lighting technology of modern vehicles. Firstly headlights with integrated daytime running lights ensure that the vehicle is recognized earlier and more clearly by other road users including pedestrians. Secondly with adaptive cornering lights, the illumination is distributed in line with the steering wheel angle and thus more efficient. This function further is enhanced by the integration of navigation data. This means that at certain road geometries, for example, at junctions, more targeted lighting is activated, which supports the driver in recognizing road users, especially those that are not well illuminated (► [Figs. 30.29](#) and ► [30.30](#)).



■ Fig. 30.29
Traffic scenario without intersection light (Berlitz et al. [2010](#))



■ Fig. 30.30
Traffic scenario with intersection light (Berlitz et al. [2010](#))

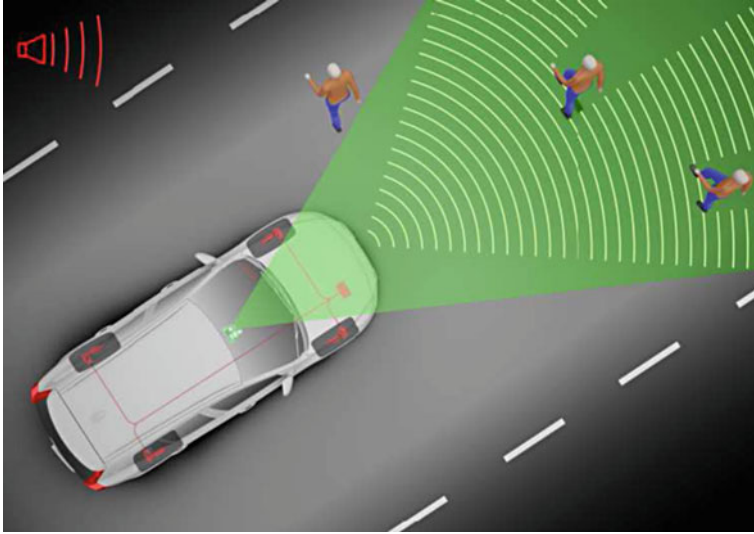
5.2.3 Autonomous System Strategies

The second main strategy for active pedestrian protection is equipping the vehicle with actuating elements that are triggered autonomously and without activation by the driver. Particularly suitable functions are an intervention in the longitudinal dynamics of the vehicle such as initiating an emergency braking. Interventions in the lateral dynamics, for example, automatic steering, are still in the research phase and are not expected to be used in production run in the near future. However a preconditioned steering by the predictive safety system which still has to be operated by the driver is imaginable. In comparison to a warning strategy whereby the reaction time of the driver must be taken into account, actions that are triggered at a relatively short time before collision have lower requirements regarding the sensor range. The approach of an algorithm which is calculating whether the vehicle and the pedestrian are following trajectories which make a collision unavoidable has already been explained in [Sect. 4.3](#).

Currently the acceptable trigger time of an automatic emergency stop falls within a relatively short time before a collision. The reason for this is illustrated in [Fig. 30.26](#). The diagram shows that if systems are activated a relatively short time before collision, a relatively low false positive rate can be expected. The increase in trigger reliability therefore conflicts with the available time-actuating elements can be effective in order to avoid or mitigate collision situations. Because of this correlation, there are high requirements on fast reaction and pressure buildup times when autonomous braking systems are used. Only in this way the relatively short time to collision can be used as well as possible ([Fig. 30.20](#)).

Combining warning systems with autonomous system intervention are also possible. For example, this includes a warning being issued to the driver with simultaneous preconditioning of the braking system. If the driver reacts to the warning, full braking power is available. The advantage of this trigger strategy is that the driver also verifies the situation for its criticality. If the driver does not react, it is still possible to intervene in the driving dynamics autonomously at a later point in time. System functionalities as this therefore combine the warning and autonomous system strategy whereby the advantages and disadvantages of both approaches also have to be taken into account ([Sect. 5.2.1](#)).

A pedestrian collision warning system including automatic emergency stop function from the manufacturer Volvo is already available on the market. It is based on sensor information from a camera and a radar sensor ([Fig. 30.31](#)). These two types of sensors are described in [Sect. 4.2](#). In the event of an upcoming collision, the driver is first warned by an acoustic and visual signal on the head-up display on the windscreen ([Fig. 30.23](#)). If the driver does not react to this warning and the collision is also unavoidable, the brakes are activated with 100% braking power. The brakes are activated less than 1 s before the calculated collision. The system can detect pedestrians and it automatically prevents collisions at speeds of up to 35 km/h. At higher speeds, the velocity of the vehicle is reduced as much as possible before impact (Volvo 2010a).



■ Fig. 30.31

Pedestrian detection with full auto brake consists of a radar unit integrated into the car's grille, a camera fitted in front of the interior rear-view mirror and a central control unit. The radar's task is to detect objects in front of the car and to determine the distance to them. The camera determines what type of object it is. In an emergency situation, the driver receives an audible warning combined with a flashing light in the windscreen's head-up display. At the same time, the car's brakes are pre-charged. If the driver does not react to the warning and an accident is imminent, full braking power is automatically applied (Volvo 2010b)

5.3 Testing of Active Pedestrian Safety Systems

For the development of active pedestrian safety systems also new methods and tools for system testing are required. This includes virtual testing in the simulation as well as testing in the real world (Keck et al. 2010). The system assessment on a pedestrian protection test facility will be explained in more detail in the following section. This test rig enables an evaluation of the complete system functionality nondestructively in close to reality situations. The vehicle-in-the-loop procedure as second test method explained combines the advantages of testing in a real vehicle with the benefits of a virtual system validation in simulation.

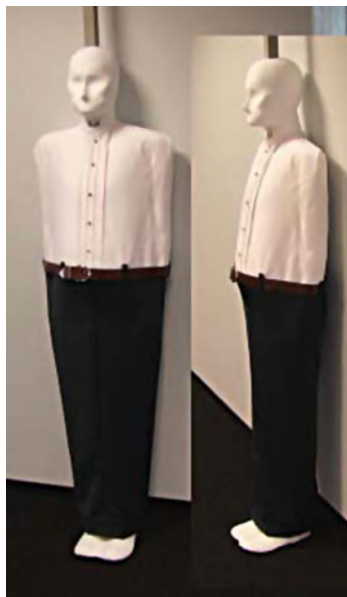
5.3.1 System Test on a Pedestrian Protection Test Facility

For safe and reproducible testing of active pedestrian protection safety systems in real-world situations, new procedures are required. The test setup shown in Fig. 30.32 can be used for this purpose. It contains of a bridge which pulls pedestrian dummies (Fig. 30.33) and other objects across the road and at short time before collision upward



■ Fig. 30.32

Pedestrian protection test facility that pulls a pedestrian dummy across the road and at short time before collision upward out of the collision area (Riedel 2008)



■ Fig. 30.33

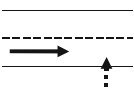
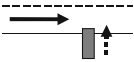
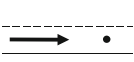
Example of a dummy for the pedestrian test facility

out of the collision area. Starting the dummy's movement and pulling it out of the risk zone is carried out on the basis of light barriers. The test facility can be used for straight road driving as well as to simulate situations where the vehicle is turning off. Most of the facility is made out of a special composite material which minimizes radar visibility to a large extent. These properties ensure that the sensors are not impaired by the rig. The dummies which are used feature specific properties depending on the sensor technology to be tested. For radar sensor analysis, the dummies can be equipped with defined clothes to ensure that the radar cross section corresponds to a human being. The contrast and the infrared reflectivity can also be controlled by modifying the clothing. So it is possible to use this facility to investigate a wide variety of sensors for their suitability as component for an active pedestrian safety function.

An examination of the findings from the accident analysis in [Fig. 30.2](#) clearly shows that this test facility can be used to investigate system behavior in realistic accident situations. As the vehicle and the pedestrian velocities can be identified from the GiDAS accident data, a catalog of realistic test cases can be drawn up from these findings, whereby the combinations of velocities reflect those that actually occur in real-world accidents ([Fig. 30.34](#)). The findings from the accident analysis can also be used to define the dummies' sizes as such information is provided by the in-depth database for the injured pedestrians.

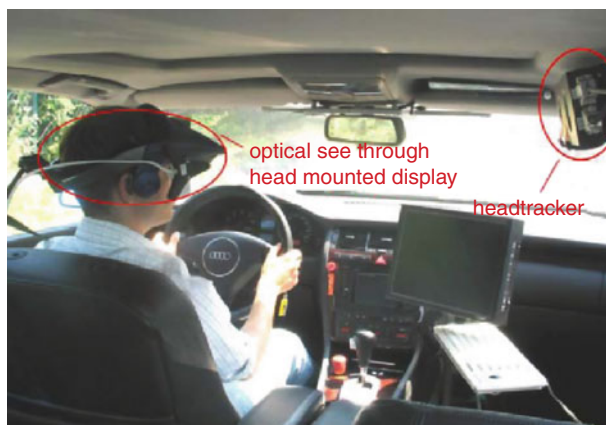
5.3.2 System Test with the Vehicle-in-the-Loop

The vehicle-in-the-loop (ViL) test procedure and simulation environment (Bock [2008](#)) combines the advantages of a real-world test vehicle with the safety and reproducibility of driving simulators. The ViL approach is based on linking the real test vehicle to a virtual traffic environment in order to benefit from the advantages of both methods. The virtual

test setup	test velocity	
	v_{vehicle}	20/40/55 km/h
	$v_{\text{pedestrian}}$	5/10 km/h
	v_{vehicle}	20/40/55 km/h
	$v_{\text{pedestrian}}$	5/10 km/h
	v_{vehicle}	20/40/55 km/h
	$v_{\text{pedestrian}}$	0 km/h

■ Fig. 30.34

Examples of pedestrian test cases that are derived from real-world collision parameters



■ Fig. 30.35

Head mounted display and head tracker in the vehicle as parts of the vehicle-in-the-loop test procedure (Bock 2008)



■ Fig. 30.36

Driver's view via the head mounted display in the augmented reality mode (Bock 2008)

traffic is seen by the driver via an optical see through head mounted display (HMD) (Fig. 30.35).

The ViL can be operated in augmented reality mode as one option. In this case, the driver still sees the real environment (e.g., lanes, roadworks) via the HMD and the traffic is faded realistic and contact analogue (Fig. 30.36). It is also possible to use the ViL in virtual reality mode (Laschinsky et al. 2010). In this case the driver can no longer see the real environment through the HMD. In fact he drives in the real car but in a virtual



■ Fig. 30.37

Driver's view via the head mounted display in the virtual reality mode

reality. The driver's view in virtual reality mode is shown in ► Fig. 30.37. The ViL test method therefore enables testing of driver assistance systems in a vehicle that is not driving in actual traffic but in open areas or on roads closed to normal traffic, for example, on a test site. The visual image of the road users or of the entire virtual environment is always restricted to the natural field of vision of the driver. As this changes depending on the position of the driver's head, the position of the HMD is constantly monitored by a head tracker (► Fig. 30.35) and the driver is only shown a certain part of the traffic simulation.

In comparison to traditional driving simulators, the ViL test method has significant advantages. For example, during the road test the driver is driving a real vehicle and therefore his awareness of risks remains unchanged. Furthermore the driver always experiences a realistic driving feeling and thus the occurrence of simulator sickness is significantly reduced. As a real vehicle is used highly dynamic road tests can also be performed with the ViL. One application of the ViL test method for active pedestrian protection is examining various warning actuating elements (Roth et al. 2009). In order to do this probands drive through a simulated city scenario in the virtual reality mode of the ViL. At defined points of the ride, pedestrians step onto the road and the test drivers have to react to these situations in order to prevent a collision. The test drivers are supported by various warnings like shown in ► Fig. 30.36. The type and duration of the reaction to the warning are analyzed in the subsequent stages. In this way, various warning strategies have been investigated with respect to their suitability for active pedestrian protection systems.

5.4 Real-World Benefit Calculation of Active Pedestrian Safety Systems

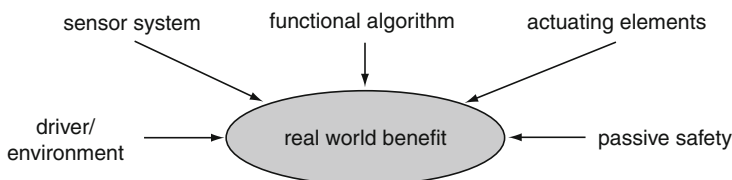
The benefit calculation which is part of the integrated development process (► Fig. 30.24) assesses the functionalities developed during the function definition stage regarding their level of protection with respect to accident avoidance and mitigation on the basis of real-world accident data. To reliably assign the positive effects produced by the active systems in real-world accidents, that is, in terms of reduction of injuries, new calculation procedures are required. The results of the benefit calculation are used to optimize the functional concepts in the process step of function definition. In this way, the active pedestrian protection system can be designed in accordance with its real-world benefit.

The real-world benefit is dependent upon several factors that are presented in ► Fig. 30.38. These include the environment sensor systems, functional algorithms, and actuating elements. The passive pedestrian protection aspects of the vehicle design the environment in which the system is used and the driver also has an influence on the field effectiveness.

5.4.1 Procedures for Calculating the Real-World Benefit

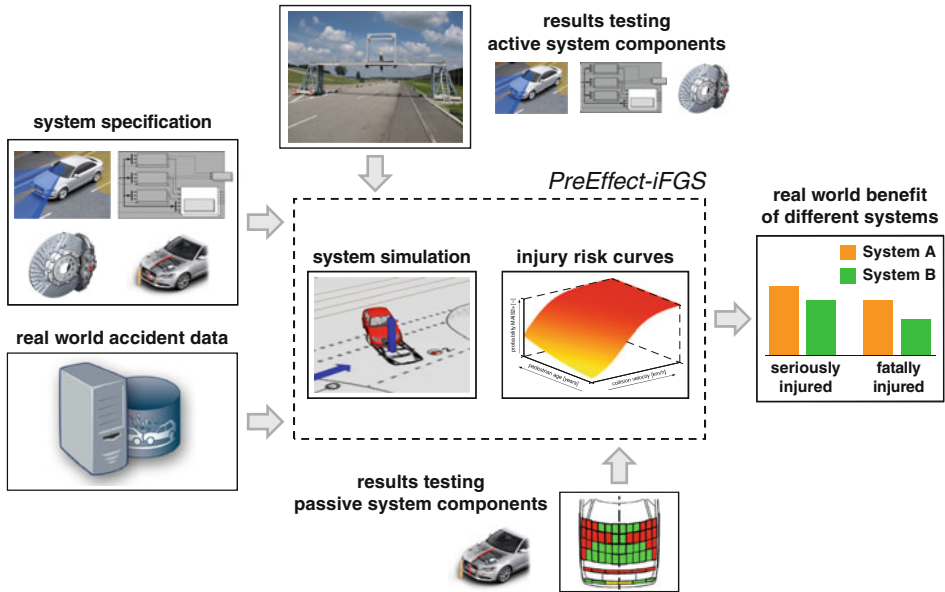
First approaches for assessing active pedestrian protection measures, for example, are discussed in Hannawald and Kauer (2003). In this study, the safety benefit of a brake assist system with respect to pedestrian collisions is evaluated on the basis of analytical calculations. A procedure for modeling just passive pedestrian safety measures on vehicles is carried out by Liers (2009). This work presents an option for assessing EuroNCAP pedestrian protection test results in real-world accidents.

The PreEffect-iFGS benefit calculation procedure developed by Schramm (2011) is a method which can be used for assessing the field effectiveness of integrated pedestrian protection systems in real-world accidents. Its process steps are illustrated in ► Fig. 30.39. In the first step, real-world accident data from the project GiDAS is imported into a software environment so that the original collision situations are available as simulation scenarios. In-depth databases are required due to the high quality of accident documentation. For example, information about the precrash phase, the site of impact at the



■ Fig. 30.38

Influencing factors on the real-world benefit of integrated pedestrian safety systems (Schramm and Roth 2009)



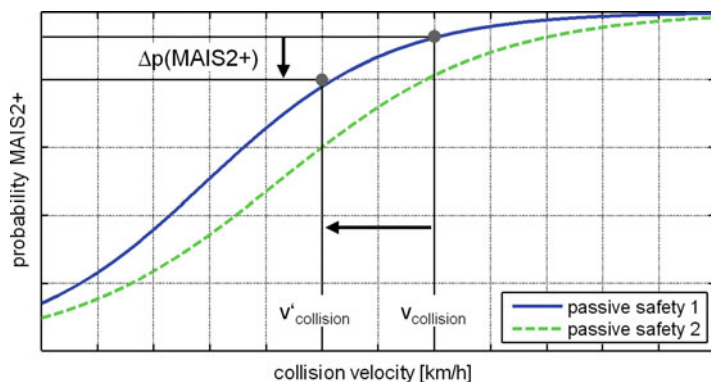
■ Fig. 30.39

Calculation procedure PreEffect-iFGS to evaluate the real-world benefit of integrated pedestrian safety systems (Schramm 2011)

vehicle, or the direction from which the pedestrian approached is needed. The GiDAS accident database (➤ Sect. 2.2.2) describes the accidents with the required amount of accuracy. In this context, one must bear in mind that the demands for standardized in-depth accident investigation and simulation scenarios for effectiveness analyses and function development will increase. In future it is therefore necessary that in-depth accident databases also provide standardized simulation scenarios of the collisions in addition to the accident parameters. This will ensure that active safety systems can be assessed on the basis of a consistent database.

In the next stage, the effect of an active pedestrian protection system is determined for each individual accident in the simulation database. This includes, for example, a reduction of the collision velocity. In order to be able to calculate the real-world benefit, PreEffect-iFGS enables the integration of test results both for the active and passive subsystem components. Taking into account the test findings for the sensor systems, functional algorithms and actuating elements is leading to the integration of real trigger times at a pedestrian test facility (➤ Fig. 30.32) instead of the functional algorithm trigger times calculated in the simulation environment. It is also conceivable to integrate reductions in vehicle speeds from tests on the pedestrian test rig. The integration of real-world findings is possible as the findings from the accident analysis are used to define the test cases. The test results can therefore also be incorporated into the real-world accident database as well (Roth et al. 2009).

The passive vehicle safety for pedestrians is taken into account in the injury risk curves (➤ Sect. 5.4.2), which are used to calculate the new injury severities resulting from the




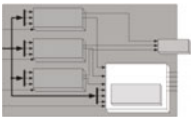

■ Fig. 30.40

Injury risk curve for pedestrians for the probability to suffer a MAIS2+ injury for two modeled passive pedestrian protection measures in relation to the collision velocity (Schramm and Roth 2009)

changed collision situations. Here the effect is that in general an improved passive protection generally leads to less severe injuries for the same impact scenario. In [Fig. 30.40](#), two injury risk curves are shown for two modeled passive vehicle feature combinations, whereby the passive safety measures 2 have a higher potential to protect the pedestrian than the passive safety measures 1. For modeling the passive safety measures, the injury shift method in accordance to Bamberg and Zellmer (1994) and Hannawald and Kauer (2003) is used in the appliance of Liers (2009). This method is necessary as there are not enough vehicles with passive pedestrian protection features in the GiDAS database to enable statistically significant injury risk curves to be created from corresponding vehicles in real accidents.

As mentioned above, the effect of passive safety systems is illustrated using injury risk curves. For example, active safety systems can reduce the collision velocity at which the vehicle hits the pedestrian ($V_{\text{collision}} \rightarrow V'_{\text{collision}}$). The extent of the reduction and its interpretation in an appropriate injury risk curve can be used to draw conclusions about the reduction of injury severity, for example, measured in the degree of MAIS2+ ([Fig. 30.40](#)). By modeling different passive measures in the injury risk curves, this process can also be carried out on the basis of different passive measures.

PreEffect-iFGS therefore enables the calculation of field effectiveness in terms of the reduction in pedestrian injury severities for different active and passive system specifications, whereby determining the changed collision situation can be based on simulation results alone as well as on findings from real-world testing. Due to the way the system components are modeled in PreEffect-iFGS, it is possible to perform a system evaluation beginning with the initial ideas and ending with the fully developed and tested safety system ([Fig. 30.41](#)). The assessment method can therefore be used across all development phases and it enables active pedestrian protection systems to be designed in accordance with their field effectiveness.

development phases			
planning	ideal	ideal	ideal
pre-development	model	ideal	model
concept development	real	real	model
series development	real	real	real

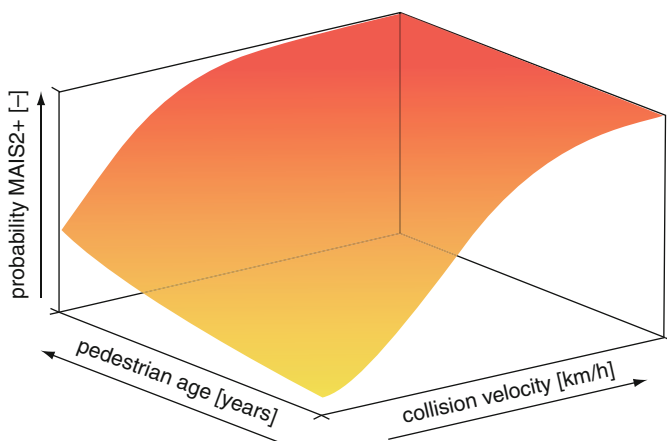
■ Fig. 30.41

Different models of the system components in PreEffect-iFGS enable a system evaluation across all development phases (Schramm and Roth 2009)

5.4.2 Generation of Injury Risk Curves for Pedestrian Collisions

As described in ► Sect. 5.4 the field effectiveness is defined as a quantifiable assessment parameter for a safety system in the context of accident avoidance and mitigation in real-world accidents. If the field effectiveness should be quantified in terms of reduction in injury severity, for example, in the number of MAIS2+ injured pedestrians, injury risk curves have to be applied which provide information on the probability that an injury will occur depending on certain factors, such as the collision velocity, pedestrian age, or impact location. Two injury risk curves just based on the collision velocity are shown in ► Fig. 30.40.

For pedestrian accidents, the collision velocity is the main determining factor for the injury severity. Furthermore the age of the pedestrian also has a significant influence on the severity of the injuries (Schramm 2011). An example injury risk curve for pedestrians who have been involved in a frontal collision with a passenger car depending on the collision velocity and the pedestrian age is shown in ► Fig. 30.42. This curve is based on the original accident data from the GiDAS database (► Sect. 2.2.1). The graph clearly shows that the injury risk rises with increasing collision velocity and increasing pedestrian age. For modeling the risk curve in ► Fig. 30.42, the multiple binary logistic regression analysis was used. As this is a parametric procedure in the proceedings of Reßle et al. (2010), Schramm (2011) also non-parametric estimation methods like k-nearest neighbor estimation and generalized additive models have been analyzed. These researches have shown that the multiple binary logistic regression analysis is an excellent method to model injury risk curves for accidents between passenger cars and pedestrians in order to calculate the probability for the pedestrian suffering a MAIS2+ injury.

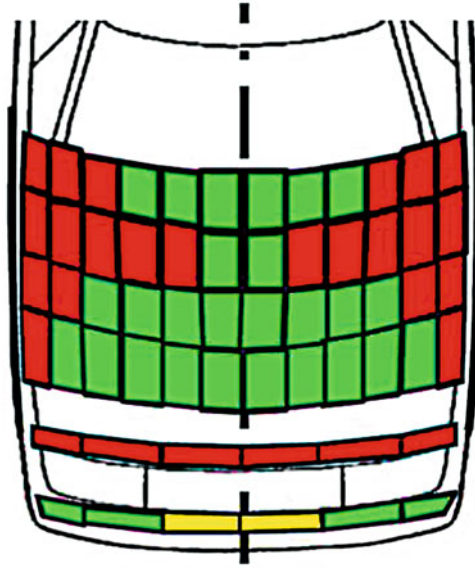


■ Fig. 30.42

Injury risk curve for pedestrians for the probability to suffer a MAIS2+ injury in frontal accidents with passenger cars depending on collision velocity and pedestrian age based on the original GiDAS pedestrian accident data

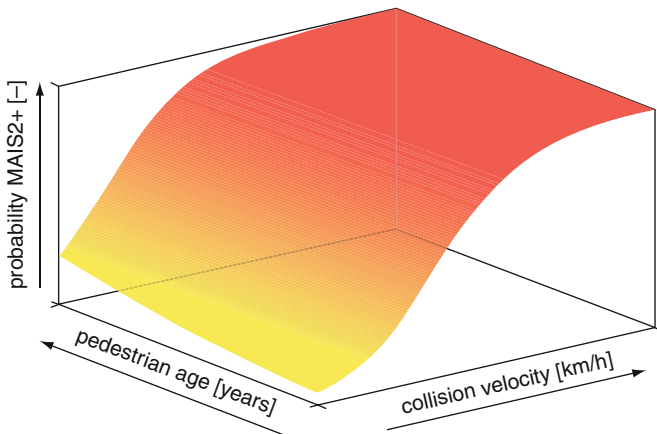
By selecting the data used to create the injury risk curves different passive pedestrian protection measures can be taken into account. As the GiDAS accident database only contains data on relatively few vehicles that have a front end optimized for pedestrian protection, statistically significant risk curves cannot be generated on the basis of these accidents. Since the PreEffect-iFGS assessment method is also intended to enable analysis of the field effectiveness of future passive strategies for pedestrian protection the accident data has to be modified. This derives from the fact that a vehicle equipment of future passive safety measures never will be found in the database. In subsequent stages, this modified accident data is then used as the basis for creating the risk curves by the multiple binary logistic regression analyses.

The injury shift method in appliance described by Liers (2009) is used to create the modified accident data which acts as the basis for generating injury risk curves. Applying this method it is possible to model the effects of purely passive pedestrian safety measures. The procedure is based on the assumption that the individual injuries to the pedestrians involved in the accidents are reduced depending on the point of impact at the vehicle. The EuroNCAP pedestrian test zones (EuroNCAP 2009) are identified for each vehicle model within the database and the individual injuries are assigned to the corresponding fields. Injuries that fall within the test area are reduced or remain unchanged depending on the test results (green, yellow, red) of a defined vehicle test result (► Fig. 30.43). The injuries are shifted on the level of AIS values (AAAM 2005), resulting in a reduced or unchanged MAIS value for the pedestrian. These new MAIS values form the basis for modeling the injury risk curves. In this way, the injury risk curves contain the model of the corresponding passive vehicle safety.



■ Fig. 30.43

Example for a EuroNCAP vehicle test result of 18 assessment points (Liers 2009)



■ Fig. 30.44

Injury risk curve for pedestrians for the probability to suffer a MAIS2+ injury in frontal accidents with passenger cars depending on collision velocity and pedestrian age for the modeled EuroNCAP test result of 18 points from ► Fig. 30.43

For a detailed description of the injury shift method to model the benefit of different EuroNCAP test results in real-world pedestrian accidents, it is referred to Liers (2009). To illustrate the principles of this method and the resulting injury risk curves, ► Fig. 30.43 shows a fictive EuroNCAP test result of 18 assessment points. Applying the injury shift

method with this test results to the real-world pedestrian accidents new MAIS values are determined for the injured persons. These MAIS values are the basis for creating an injury risk curve for a modeled passive protection level of 18 Euro NCAP points according to [Fig. 30.43](#). A comparison of the risk curves in [Fig. 30.43](#) which is based on the unchanged MAIS values from GiDAS database and [Fig. 30.44](#) which is based in the new less or equal MAIS values compared to original GiDAS database shows that the effect of an improved passive protection level is reflected in a decreasing injury probability.

6 Conclusion

Pedestrian protection is an important field of today's vehicle safety. In the past especially infrastructural measures or passive pedestrian vehicle safety contributed in reducing pedestrian casualties. In the future, more and more active safety technologies for pedestrian protection will be found in modern vehicles. The development of these innovative safety systems requires new processes and tools. At the beginning of the development, it is necessary to analyze the pedestrian accidents. For this step in-depth accident databases have to be taken into account because currently only these databases contain highly detailed accident documentations. This information can be used for creating simulation databases representing real-world collision situations. Based on these defined components or the total system basically consisting of environment sensor systems, functional algorithms and actuating elements are analyzed. In this context, two main system strategies have to be distinguished. On the one hand strategies that autonomously engage into the driving situation and on the other hand strategies which draw the driver's attention to a dangerous situation by presenting a warning. Besides the simulation also real-world testing has to be accomplished, for example, on a pedestrian test rig which pulls a human dummy across the road and at short time before collision upward out of the collision area. For the evaluation of warning strategies, it is necessary to analyze the drivers' reaction on different warning elements. Here the application of the vehicle-in-the-loop test procedure is suitable. This method combines the real test vehicle with a virtual traffic environment. Another important aspect of developing active pedestrian safety systems is the process step of benefit calculation. In this step, the total systems are assessed regarding their level of protection with respect to accident avoidance and mitigation on the basis of real-world accident data. For determining the real-world benefit of active pedestrian safety systems, the calculation procedure PreEffect-iFGS is described in greater detail which enables a system evaluation from the initial ideas up to the fully developed and tested safety system. In the future also other vulnerable road users like bicyclists will be focused by the field of active vehicle safety whereupon today's active pedestrian safety systems can already have positive effects on these other vulnerable road users. The presented development process and methods for designing, testing, and calculating the real-world benefit can be used as basis for these future challenges. Further the communication between the different

road users, for example, by car-to-X technologies, can also contribute in reducing the number of pedestrian or other road user casualties in the near future.

References

- AAAM (2005) Abbreviated injury scale 2005. Association for the Advancement of Automotive Medicine, Barrington
- Bamberg R, Zellmer H (1994) Nutzen durch fahrzeugseitigen Fußgängerschutz. Berichte der Bundesanstalt für Straßenwesen, Fahrzeugtechnik, F5
- Berlitz S, Funk C, Kenn C (2010) Lichtassistentensysteme und adaptive Beleuchtung im Audi A8. Elektronik automotive Sonderausgabe Audi A8:14–15
- Bock T (2008) Vehicle in the loop – test- und Simulationsumgebung für Fahrerassistenzsysteme. Dissertation TU München, Cuvillier, Göttingen
- Botsch M, Lauer C (2010) Complexity reduction using the random forest classifier in a collision detection algorithm. Intelligent vehicles symposium, San Diego
- ECE (2009) Global technical regulation no. 9- Pedestrian safety. <http://www.unece.org/trans/main/wp29/wp29wgs/wp29gen/wp29registry/gtr9.html>. Accessed 10 Nov 2010
- EuroNCAP (2009) Pedestrian testing protocol. <http://www.euroncap.com/files/Euro-NCAP-Pedestrian-Protocol-Version-5.1-0-7caffdd8-98e2-4b6d-a808-8d9a4341930.pdf>. Accessed 10 Nov 2010
- European Parliament (2009) Regulation (EC) No 78/2009 of the European parliament and of the council of 14 January 2009
- European Commission (2007) Impact assessment – proposal for a regulation of the European parliament and of the council on the protection of pedestrians and other vulnerable road users. SEC(2007) 1244
- FGSV (2002) Empfehlungen für Fußgängerverkehrsanlagen. FGSV Nr. 288, Köln
- GiDAS (2010) German in-depth accident study. <http://www.gidas.org/>. Accessed 10 Nov 2010
- Hannawald L, Kauer F (2003) Equal effectiveness study on pedestrian safety. Technische Universität Dresden
- Heißing B, Ersoy M (2008) Fahrwerkhandbuch – Grundlagen, Fahrdynamik, Komponenten, Systeme, Mechatronik, Perspektiven. Vieweg+Teubner, Wiesbaden
- Hoffmann J, Gayko J (2009) Fahrerwarnelemente. In: Winner H, Hakuli S, Wolf G (eds) Handbuch Fahrerassistenzsysteme. Vieweg+Teubner, Wiesbaden, p 345
- IRTAD (2010) IRTAD Database, June 2010 – Fatalities by road use. <http://internationaltransportforum.org/irtad/pdf/roaduse.pdf>. Accessed 10 Nov 2010
- Keck F, Kuhn A, Sigl S, Altenbuchner M, Palau T, Roth F, Stoll J, Zobel R, Kohsiek A, Zander A (2010) Prüf- und Evaluationsverfahren für den vorausschauenden Fußgängerschutz im Spannungsfeld zwischen Simulation und realer Erprobung. 15. VDI-Tagung – Erprobung und Simulation in der Fahrzeugentwicklung, Baden Baden
- Laschinsky Y, von Neumann-Cosel K, Gonter M, Wegwerth C, Dubitzky R, Knoll A (2010) Evaluation of an active safety light using virtual test drive within vehicle in the loop. IEEE-ICIT, Viña del Mar – Valparaíso, Chile
- Liers H (2009) Benefit estimation of the Euro NCAP pedestrian rating concerning real world pedestrian safety. 21. In: Conference on the enhanced safety of vehicles, Stuttgart
- Maier R (1984) Fußgängersicherheit in Städten. Dissertation Universität Karlsruhe, Mitteilungen der Beratungsstelle für Schadenverhütung, Nr. 24
- PMDTec (2010) PMDTechnologies GmbH
- Reßle A, Schramm S, Kölzow T (2010) Pedestrian injury risk functions for real world benefit evaluation of integrated safety systems. Crash.tech, Leipzig
- Riedel H (2008) Trends aktiver Sicherheitssysteme von Fahrzeugen. Technologien in Bewegung – Kunststoffe, neue Prozesse und Produkte, Traboch
- Roth F, Maier K, Stoll J, Dubitzky R, Zander A, Schramm S (2009) Integraler Fußgängerschutz – Funktionsentwicklung. 4. Praxiskonferenz Fußgängerschutz, Bergisch-Gladbach
- Schramm S (2011) Methode zur Berechnung der Feldefektivität integraler Fußgängerschutzsysteme. Dissertation TU München

- Schramm S, Roth F (2009) Method to assess the effectiveness of active pedestrian protection safety systems. 21. In: International technical conference on the enhanced safety of vehicles, Stuttgart
- Taner A, Rosenow A (2010) Nachtsichtassistent. Elektronik automotive Sonderausgabe Audi A8:96–98
- Volvo (2010a) Kollisionswarnsystem mit Fußgängererkennung und automatischer Notbremsung. <http://www.volvocars.com/at/explore/Pages/pedestrian-detection.aspx>. Accessed 10 Nov 2010
- Volvo (2010b) Volvo car corporation – global newsroom. <https://www.media.volvocars.com/de/>. Accessed 10 Nov 2010
- WHO (2009) Global status report on road safety – time for action. http://whqlibdoc.who.int/publications/2009/9789241563840_eng.pdf, p 240. Accessed 10 Nov 2010

31 Parking Assist

Michael Seiter¹ · Hans-Jörg Mathony¹ · Peter Knoll²

¹Robert Bosch GmbH, Leonberg, Germany

²Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

1	<i>Introduction</i>	830
2	<i>Ultrasonic Technology for Automotive Sensors</i>	831
2.1	Fundamentals of Ultrasonic Technology	831
2.2	Piezoelectric Effect and Piezoelectric Materials	831
2.3	Emission of the Ferroelectric Transducer	832
2.4	Detection Area of Ultrasonic Transducers	833
2.5	Distance Measurement	833
2.6	Ultrasonic Sensor	835
2.7	Sensor Integration into the Bumper	838
3	<i>Video Technology for Automotive Application</i>	839
3.1	Fundamentals of Photosensing	839
3.2	CCD Sensors	840
3.3	CMOS Sensors and Imagers	841
3.4	Structure of Video Cameras and of a Vision System	843
4	<i>Parking and Maneuvering Assistance Systems</i>	844
4.1	Requirements to Parking Assistance Systems	844
4.2	Categories of Parking and Maneuvering Assistance Systems	845
4.2.1	Passive Systems	846
4.2.2	Active Systems	857
5	<i>Benefit of Parking Systems</i>	859
6	<i>Functional Limits of Parking Assistance Systems</i>	860
7	<i>Legal Aspects of Parking Assistance Systems</i>	861
8	<i>Conclusion</i>	862

Abstract: On virtually all motor vehicles, the bodies have been designed and developed in such a way as to achieve the lowest possible drag coefficient values in order to reduce fuel consumption. This trend has resulted in a gentle wedge shape which greatly restricts the driver's view when maneuvering. Obstacles can only be poorly discerned – if at all.

To overcome these problems, ultrasonic-based parking aids were introduced in the European market in the early 1990s. These systems monitor the rear and the front of the vehicle, and warn the driver if there is an obstacle which can cause a collision. Recently, new functions like semiautomatic Parking Assistance have been realized based on the same sensor technology. Such a system automatically steers the vehicle into the parking space while the driver controls the longitudinal movement of his car.

In parallel, cameras to monitor the rear of the vehicle have first been introduced on the Japanese market together with central information systems allowing presenting its picture in the center console area. Due to the availability of powerful image processing units, recently multi-camera systems have been launched. These systems fuse the data of four cameras, for example, to create a 360° top-view picture showing the surroundings of the vehicle.

Further sensor improvement and system development of both ultrasonic and camera technology as well as sensor data fusion of different technologies will allow new parking and maneuvering functions with increasing automation grade.

The chapter starts with basics of ultrasonic and camera technology. Furthermore, it emphasizes on driver assistance systems for parking and for slow maneuvers based on these sensor technologies.

1 Introduction

Parking assistance and maneuvering systems are comfort functions. This means that the driver is responsible for his vehicle at all times. The first systems coming on the market were called “Parking aids” or “Park Distance Control.” They monitored the front and rear of the vehicle and warned the driver if there was an obstacle in the vicinity of the vehicle when maneuvering, entering or leaving a parking space. Ultrasonic sensors are widely used for these systems.

Once the vehicle is equipped with these sensors for the parking aid, other functions like Parking Assistance and Maneuvering systems can be realized. They have been introduced step-wise. The first system, the parking-space measurement system, measures the length of a parking space and gives the driver information whether the parking space is large enough for the vehicle or not. In the next step, recommendations how to turn the steering wheel are given to the driver to enter best into the parking space. If the vehicle is equipped with an electromechanical steering, the systems steers the vehicle into the parking space, while the driver must care only for the longitudinal control of his vehicle. Fully automatic systems that drive (steer and brake) the vehicle automatically into the parking space have been shown as prototypes and are in particular attractive for small garages or narrow parking spaces.

Cameras to monitor the rear of the vehicle are used since more than 10 years, mainly in Japan. The picture from the camera which is placed near the rear number plate of the vehicle is shown on a graphic screen, placed in the center console region of the car. Often the picture distortions caused by the wide-angle characteristic of the camera lens is corrected by a microprocessor. Auxiliary lines, e.g., the vehicle's course depending on the steering angle or distance lines can be added to the picture to help the driver to estimate better the scenery behind the car while reversing. Combining a camera with an above mentioned ultrasonic parking aid system gives the driver valuable information about the distance to objects displayed on the screen. Meanwhile, systems have been introduced on the market with four cameras in front, in the rear and on the two sides of the car. By means of picture processing, the pictures of the four cameras are fused, and the system gives not only a complete around view, but the driver can also choose different views including a bird's-eye view on the own car. This feature helps significantly in a docking situation. Information extraction from the camera pictures and picture processing, as well as sensor data fusion with the ultrasonic sensor system will allow new parking and maneuvering functions with continuously increasing automation grade.

From a technical point of view, fully autonomous parking will be possible with the driver using a remote control from outside the vehicle to park his car in any parking space which is large enough. To realize these kinds of systems, however, also legal requirements and aspects have to be addressed accordingly.

2 Ultrasonic Technology for Automotive Sensors

2.1 Fundamentals of Ultrasonic Technology

Ultrasonic sensors are used for many applications. Besides automotive application in the past 20 years, they are used since long time for military application in submarines, in Medicine for diagnostics, and as sensors for distance measurement in industrial applications. The physical basics can be found in the literature, e.g., (Waanders 1991) together with the description of different applications.

Piezoceramic ultrasonic transducers are small and very robust and therefore ideal for automotive applications. They are in use since 1993 in series applications and will be described in detail in the first chapter.

2.2 Piezoelectric Effect and Piezoelectric Materials

The piezoelectric effect describes the electromechanical context between the electric and the mechanic status of a crystal. If there is an electric field applied at the electrodes on two sides of a piezoelectric crystal, a mechanical deformation results. Vice versa, a mechanical deformation of the crystal results in an electric voltage, which can be measured at the crystals electrodes. It is proportional to the deformation. The effect is very fast and, as

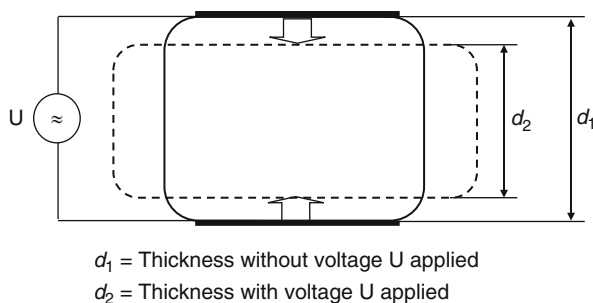
a result, piezoelectric materials can be used for the generation of oscillations with rather high frequencies and for the reception of sound waves. With other words, an ultrasonic piezo element is a “loudspeaker” and a “microphone” in one unit, and they are therefore called “transducers.” A more detailed description of the piezoelectric effect can be found in Noll and Rapps (2009).

2.3 Emission of the Ferroelectric Transducer

Ultrasonic transducers which emit sound into air and receive sound from the air must have rather high amplitudes in order to couple enough energy into the air. The mechanical deformation of the piezoceramic material alone is not sufficient, making it necessary to amplify the effect by mechanical means. In practice, this is made by gluing the ceramic material on a metallic membrane. If a voltage is applied at the electrodes, the ceramic material changes its diameter and its thickness, see Fig. 31.1. If the ferroelectric material is fixed on the metallic membrane, these changes are transferred into a bend-oscillation of the membrane which generates much larger oscillating amplitudes, especially when operated at the resonance frequency.

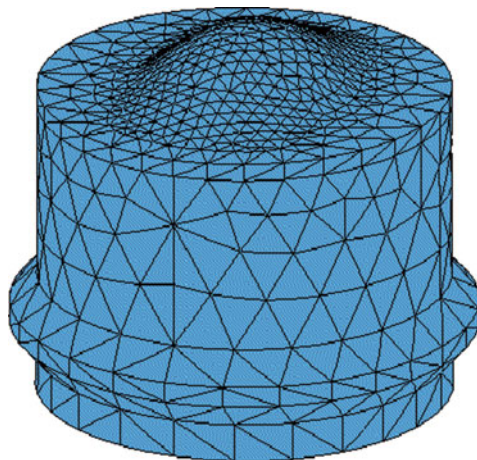
Vice versa, an incoming acoustic wave creates oscillations of the membrane and thus contributes to a change of the diameter of the ceramic material. As a consequence, an alternating voltage is generated at the electrodes which can be amplified and further processed. Mostly, ultrasonic transducers are used as transmitters as well as receivers. This is insofar from relevance as the fixation of the pot influences the movement of the membrane. Figure 31.2 shows a computer simulation of an ultrasonic transducer operated at the primary oscillation frequency. To avoid inner oscillations of the transducer, sound-absorbing foam is applied on the inner side of the transducer.

For ultrasonic transducers in automotive parking aid systems, an operating frequency between 40 and 50 kHz is commonly used. This has been proved as the best compromise between good acoustical performance (sensitivity and range) and high robustness against noise from the surrounding of the transducer (Noll and Rapps 2009).



■ Fig. 31.1

Deformation of a piezoceramic by applying a voltage U to the electrodes (Knoll 2010)



■ Fig. 31.2

FEM simulation of an oscillating membrane operated at resonance frequency (Bosch 2011)

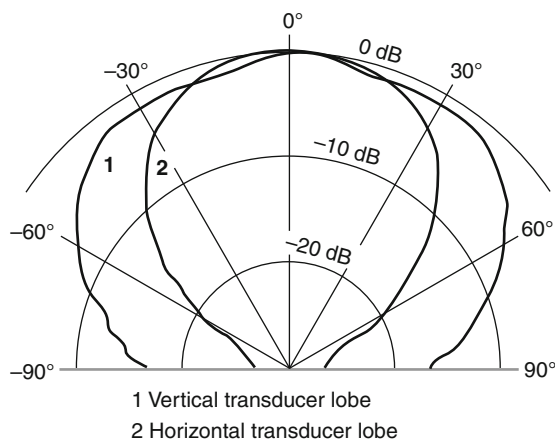
2.4 Detection Area of Ultrasonic Transducers

The detection field of an ultrasonic transducer is one of the main quality features for the resulting environment detection function. In order to detect the widest possible range, the detection characteristics must meet special requirements to allow a complete coverage of the detection area without gaps and with as few sensors as possible the detection field of the single sensor should be very wide (120° – 140°). At the same time, in the vertical range, it is necessary to have a smaller detection angle in order to avoid interference from ground reflections in particular on gravel roads (clutter). In practice, a value between 60° and 70° (about half of the horizontal angle) has shown to be advantageous for the vertical detection angle, see ► Fig. 31.3.

Short development times and diverse vehicle integration conditions for the sensors in the bumper require an efficient prediction of the expected acoustical sensor performance with regards to the location of the sensor in the bumper in an early state of the project. Mature simulation methods meanwhile allow a reliable prediction without the need to produce hardware-prototypes and to make costly tests. For the simulation of sound propagation, the Boundary Element Method, BEM, has proved to be best. With this method, only the sound-emitting surface is calculated, not the complete volume, see ► Fig. 31.4. This reduces the calculating effort significantly.

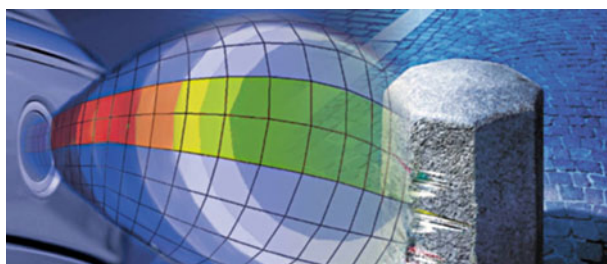
2.5 Distance Measurement

Due to the slow propagation speed of sound the distance measurement with ultrasound on the basis of the time-of-flight measurement principle of a signal is rather easy.



■ Fig. 31.3

Antenna radiation diagram of an ultrasonic sensor (Bosch 2011)



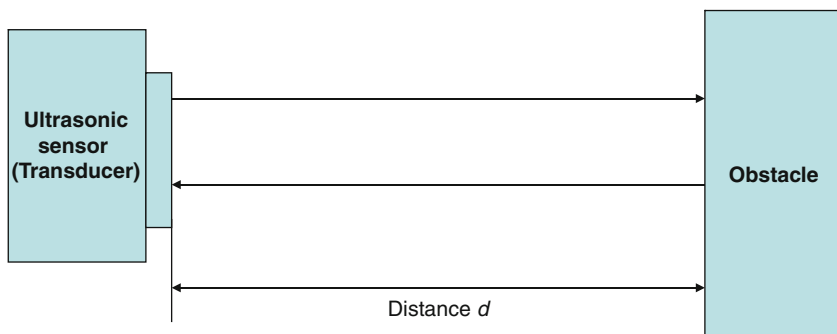
■ Fig. 31.4

Simulation of the emission lobe of an asymmetrical ultrasonic pulse ($t \sim 0, 2$ s) for the parking aid (Bosch 2011)

Following a principle that is similar to echo depth sounding, the sensors transmit ultrasonic pulses with typically about 300 μ s, at a frequency of approximately 40 kHz, and measure the time taken for the echo pulses to be reflected back from obstacles. These ultrasonic sensors operate according to the pulse/echo principle in combination with trilateration. The distance of the sensor to the nearest obstacle is calculated from the propagation time of the first echo pulse to be received back according to the equation:

$$d = 0.5 \cdot t_e \cdot c$$

with t_e : propagation time of ultrasonic signal(s), c : velocity of sound in air (approximately 340 m/s).



■ Fig. 31.5

Measurement principle of an ultrasonic transducer (Knoll 2010)

► Figure 31.5 shows the principle of ultrasonic distance measurement. Ultrasonic pulses are emitted by the transmitter, are reflected by the obstacle, and received by the receiver.

The accuracy of the measurement is influenced by a couple of factors. On one hand, these are the physical dependencies of the propagation speed of sound in air as propagation medium. Here, in particular, the air temperature is of major influence. It must be compensated by electronic means.

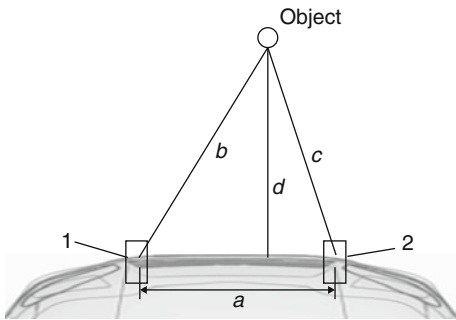
More critical are the geometric measurement inaccuracies caused by the position, the outlines, the geometry, and the orientation of obstacles relevant to the sensor. Decisive for the reduction of these influences is the use of multiple sensors on the vehicle front (typically 4–6) and, on the other hand, the application of the so-called trilateration method. A single ultrasonic sensor measures the direct distance between *sensor* and obstacle, but in practice, the distance between *vehicle* and obstacle is relevant for a warning to the driver. ► Figure 31.6 shows an example for the distance calculation by trilateration. Each sensor receives the own echo and the echo from the neighbored sensors (cross-echo) and calculates the real distance to the vehicle as a projection on the bumper with the formula shown in the figure.

► Figure 31.7 shows a typical signal structure: On the left side of the image the transmitted impulse is shown, while the echo pulse is shown in the right part of the figure (Bosch 2006, 2011).

2.6 Ultrasonic Sensor

The sensor consists of a plastic housing with integrated plug-in connection, an ultrasonic transducer, and a printed circuit board with the electronic circuitry to transmit, to receive, and to evaluate the signals, see ► Fig. 31.8.

The acoustical part of the ultrasonic sensor consists of the aluminum pot with the piezo element on the inner side. The two electrical connections from the piezo element to

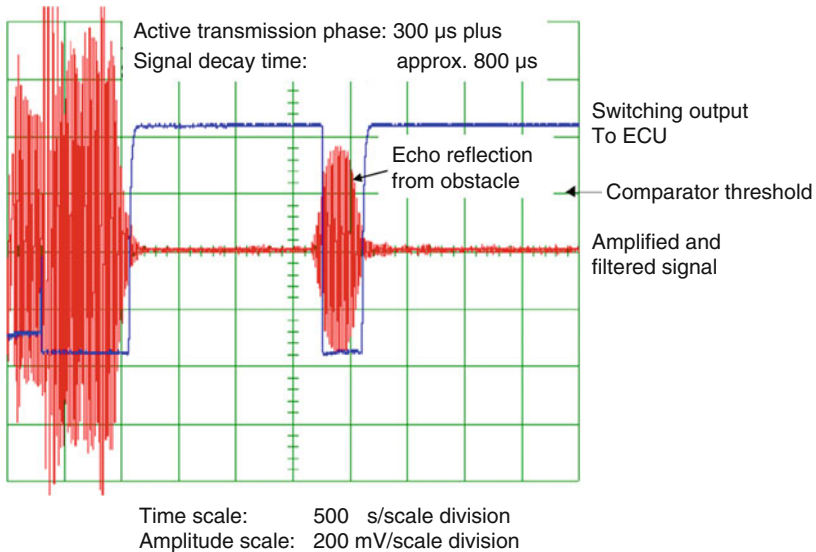


- 1: Sensor 1
 2: Sensor 2
 a: Distance sensor 1 to sensor 2
 b: Distance sensor 1 to obstacle
 c: Distance sensor 2 to obstacle
 d: Distance between bumper and obstacle

$$d = \sqrt{b^2 \frac{(a^2 - c^2 + b^2)^2}{4a^2}}$$

■ Fig. 31.6

Calculating the distance from a single obstacle by trilateration (Bosch 2011)



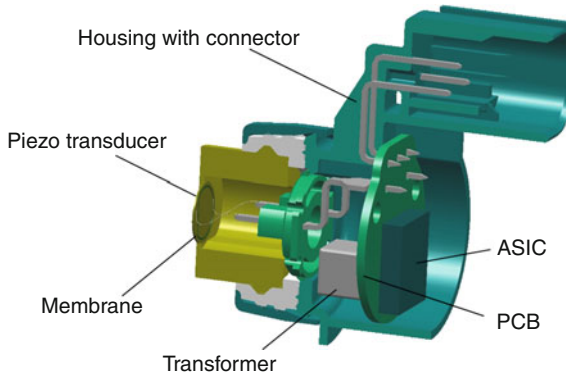
■ Fig. 31.7

Signals of an ultrasonic transducer (Bosch 2011)

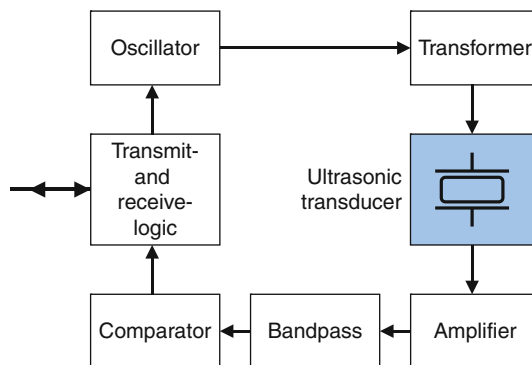
the PCB must be made with very thin and soft connectors to avoid an acoustic coupling of the PCB to the membrane.

► [Figure 31.9](#) shows a block diagram of the sensor electronics.

The sensor is electrically connected to the ECU by three wires, two of which supply the power. The third, bidirectional signal line, is responsible for activating the transmit function and for returning the received and evaluated signal to the ECU. When the sensor receives a digital transmit pulse from the ECU, the electronic circuit excites the aluminum



■ Fig. 31.8
Cross-section of an ultrasonic sensor (Bosch 2011)



■ Fig. 31.9
Block diagram of an ultrasonic sensor (Bosch 2011)

membrane with square wave pulses at the resonance frequency so that it vibrates, and ultrasound is emitted. During the time taken for it to stop oscillating (approx. 700 μ s) no reception is possible. This limits the minimum measurable distance to about 20 cm. The membrane, which has meanwhile returned to rest, is made to vibrate again by the sound reflected back from the obstacle. These vibrations are converted by the piezoceramic wafer to an analog electrical signal which is then amplified and converted to a digital signal by the sensor electronics. Sensors with digital interface calculate the distance from the time-of-flight of the signal and transmit it to the ECU of the system (Bosch 2008).

The housing of the sensor must firstly protect the transducer and the electronic circuitry from environmental influences. Secondly, it must ensure the plug-connection to the wire harness of the vehicle and, thirdly, the mechanical connection of the sensor with the bumper. Usually the sensor housing is filled with a Silicone material to protect the

sensor components from water and to avoid that undefined cavities may influence the acoustical behavior of the sensor (Bosch 2011).

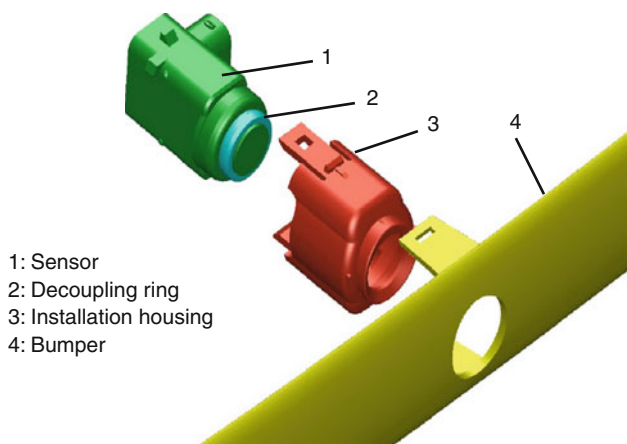
2.7 Sensor Integration into the Bumper

The design of the sensor and its fixation in the bumper must ensure that the sensor is rarely noticeable. The visible part of the sensor must be paintable in the bumper color without influencing the functionality of the sensor. When the sensor is integrated in the vehicle's bumper, the membrane of the pot is (almost) equal to the surface of the bumper and is usually painted with the same color as the bumper. Also chromium surfaces for the membrane are possible.

The latest generation of ultrasonic transducers can be better adapted to the vehicle's contours as previous generations. Resistance against vibrations, temperature changes, weather influence, resistance against humidity, and a reliable acoustical decoupling of the membrane from its surrounding are of particular importance.

Specifically adapted mounting brackets secure the sensors in their respective positions in the bumper, see [Fig. 31.10](#). To avoid crosstalk between neighbored sensors by structure borne noise, it is essential that the oscillations of the membrane are completely decoupled from the housing of the sensor. Therefore, the membrane pot is embedded into a soft Silicone ring whose acoustical properties must not change with low temperatures and with age.

The installation angle of and the distances between the sensors are measured on a vehicle-specific basis. This data are taken into account in the ECU's calculation algorithms (Noll and Rapps 2009; Bosch 2011).



■ Fig. 31.10

Installation principle of the ultrasonic sensor in the bumper (Bosch 2011)

3 Video Technology for Automotive Application

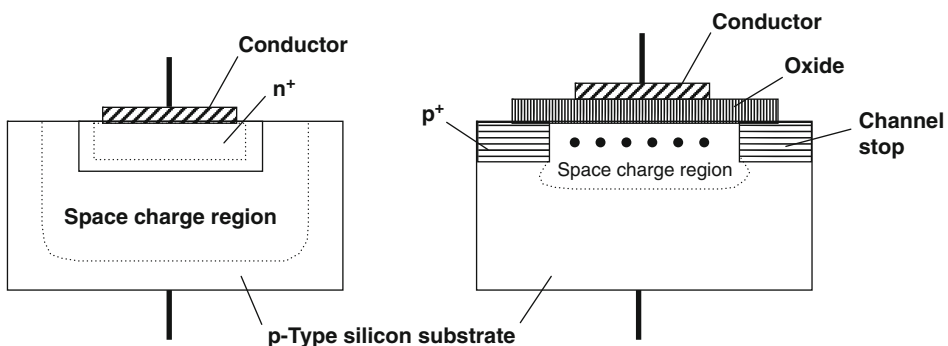
3.1 Fundamentals of Photosensing

Incident photons follow their paths into the interior of a semiconductor device, where most of the photons interact by producing electron-hole pairs. These photocharge pairs need to be separated in an electric field before they recombine again, leading to the flow of a photocurrent, which is proportional to the incident light intensity over many orders of magnitude.

❶ *Figure 31.11* illustrates the two most important photosensitive structures, the photodiode (PD, left) and the metal-oxide semiconductor (MOS, right) capacitor as used in the charge-coupled device (CCD) Image sensors. Both devices are easily fabricated with standard semiconductor processes.

A photodiode consists of a combination of two different conductivity types of semiconductor, as illustrated in ❶ *Fig. 31.11*, left. In the junction between the two types of semiconductor, an electric field exists in the space-charge region. Photodiodes are typically operated by biasing them to a certain potential and exposing them to light. Photocharge pairs entering the space-charge region are separated in the PD's electric field, a photocurrent is produced, and the photocharge is accumulated in the PD's capacitance. After the exposure time, the residual voltage is measured, and the voltage difference compared with the reset voltage level is a measure for the amount of light incident in the pixel during the exposure time.

The MOS-capacitance illustrated in ❶ *Fig. 31.11*, right, consists of a thin layer of oxide on top of a semiconductor material. The oxide is covered with a conductive material (e.g., Polysilicon). As in the case of the PD, the MOS structure is biased to a suitable voltage, leading to a space-charge region of a certain extent in the semiconductor. Again, photocharge is separated in the electric field, and it is integrated in the MOS capacitance, collected at the interface between semiconductor and oxide.



■ **Fig. 31.11**
Photosensitive structures. *Left: Photodiode, right: MOS capacitance* (Knoll 2003)

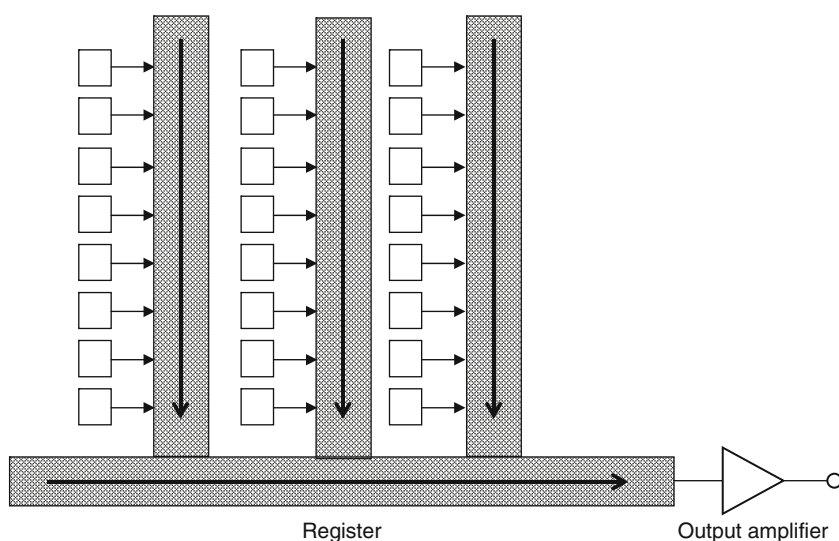
Once the storage capacity of the sensitive elements is exceeded, additional photo charge in the corresponding pixel starts to spill over to neighboring pixels. This effect is called blooming.

3.2 CCD Sensors

In the case of CCDs, the photo charge is stored under a precharged MOS capacitance. The basic CCD idea is to combine a linear array of MOS capacitances so that a stored photo charge can be moved laterally under the influence of appropriate MOS electrode voltage patterns. Photocharge pairs are generated in the semiconductor under the influence of incident light. Moving by diffusion and by drift, the photoelectrons can find their way to positively biased MOS electrodes (Gates), where they are stored at the interface between semiconductor and thin oxide. The photo-generated holes are repelled by the positive gate voltage, and they move around by diffusion until they finally combine in the silicon substrate.

Interline-transfer CCDs are the mostly used type in consumer applications. **Figure 31.12** shows the basic principle: The charges are sequentially and vertically transferred to a register.

CCD sensors suffer from a limited dynamic range. State-of-the-art rear-view cameras are still a domain of CCD technology. With the technological mainstream in the consumer area, especially with consumer camera development, it is expected that for more demanding vehicle application also, that this product domain will be dominated by CMOS technology in the future.



■ Fig. 31.12

Interline-transfer CCD with column light shields for vertical charge transfer (Knoll 2003)

3.3 CMOS Sensors and Imagers


CMOS-sensors use non-integrating photodiodes as sensor elements which are, at the same time, independent from the exposure time. The luminance signals are logarithmized before the signals are selected. This results in a characteristic similar to the human eye. Only by this means, it is possible to reach a dynamic range of 100 dB and more.

CMOS sensors feature several advantages with respect to the more generally used CCDs: they are fabricated in a fully standard VLSI technology, which means low costs by taking advantage of submicron CMOS technology. Several functionalities can be integrated on the sensor substrate, including random access. They consume very little power as the circuitry in each pixel is typically active only during the readout period, and there is no clock signal driving large capacitance. Readout speed can be enhanced by parallel access to multiple taps of the pixel array. Because of these characteristics, CMOS sensors are the favored technology for demanding application like automotive.

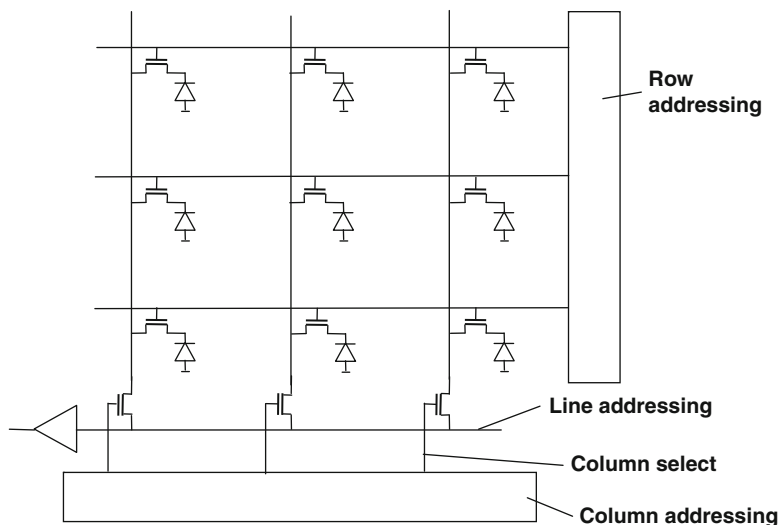
Recent improvement in CMOS process technology, pixel architecture, and sensor principles have launched a variety of imager types, color, as well as black and white solutions. Common for all developments is the pressure to minimize the diagonal of the pixel array size to take profit from the cost saving of smaller lens systems.

The human eye shows a nonlinear, close to logarithmic sensitivity, and it is obvious to realize this behavior also in a photosensor. The realization of CMOS pixels offering a logarithmic sensitivity is particularly easy to obtain: one can use the logarithmic relationship between gate voltage and drain current in a MOSFET operated in weak inversion. The resulting pixel architecture is easy to implement in a CMOS process because a pixel consists of just a Photodiode and three MOS transistors. A typical photo-response of about 40 mV per decade of optical Input Power is obtained with such logarithmic pixels, and their useful dynamic range exceeds 120 dB.

The opto-electric conversion characteristic can be designed as linear or as nonlinear one.

A CMOS Imager consists of a two-dimensional arrangement of photodiodes, each provided with its own selection transistor, as shown in  Fig. 31.13. For the description of the operation, assume that all photodiodes are precharged to a certain reverse bias voltage, typically 5 V. Under the influence of the incident light, each pixel is discharged to a certain level. A pixel is read out by addressing the corresponding row and column transistors, providing a conducting line from the pixel to the output amplifier. Using this line the Pixel is charged up again to the same reverse bias voltage as before. The amplifier measures how much charge is required to do this, and this charge is identical to the photo charge (plus dark current charge) accumulated at the pixel site. In this way, each pixel can be read out individually, at random, and the exposure time is completely under the control of the external addressing electronics.

With the so-called Active Pixel Sensor (APS) imaging technology, the noise behavior has been significantly improved by placing the first MOSFET into the pixel. The simplest



■ Fig. 31.13

Photodiode array image sensor with one photodiode and one selection transistor per pixel (Knoll 2003)

APS image sensor pixel consists of one photodiode and three MOSFETs. It is very attractive for several reasons:

- APS-pixels can be produced in standard CMOS technology, opening the way to image sensors with integrated electronic functionality and even complete digital processors
- The pixels offer random access similar to PD arrays
- The exposure times can be programmed electronically
- APS image sensors dissipate one or two magnitudes less electrical power than CCDs
- APS imagers show less blooming (spilling of electronic charge to adjacent pixels).

CMOS sensors have achieved high spatial resolution. The pixel size is usually in the region of $5\ \mu\text{m}$. Resolution or imager size or format is a technical parameter that is only limited by the size and by the desired pixel opto-sensitivity. A small pixel or a high resolution implies a low sensitivity and vice versa. Designing an imager for automotive applications means to balance a variety of trade-offs.

A high geometrical resolution (e.g., $4\text{ k} \times 4\text{ k}$) can be only achieved by squeezing pixel pitch and in sequence the fill factor and sensitivity. The need for deep submicron CMOS processes shifts the maximum of the spectral response toward the blue end of the visual spectrum with small quantum efficiency.

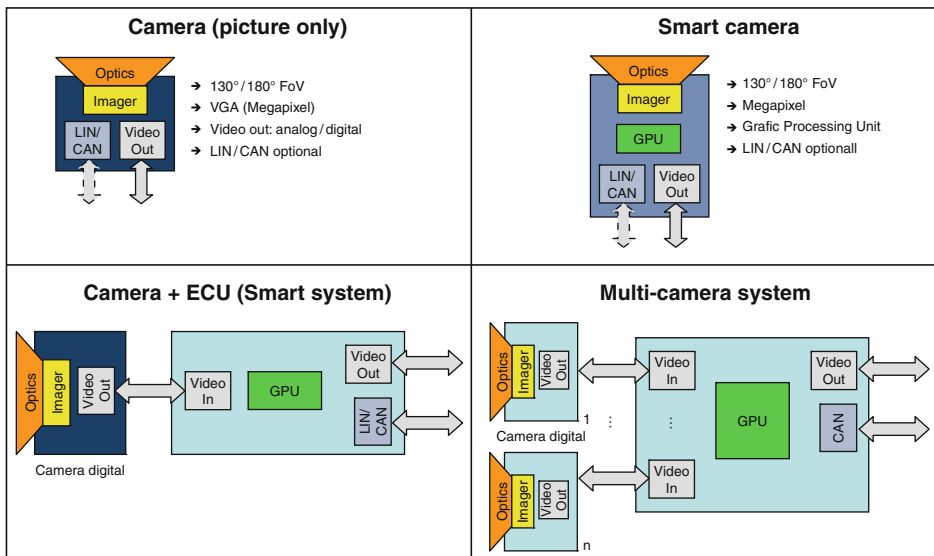
➤ *Figure 31.14* shows the difference in dynamic behavior of a CCD and a nonlinear CMOS sensor. The difference is obvious: While the CCD-sensor shows almost no details within the light area at the end of the tunnel, these details (trees, cars, road structures) can be clearly seen in the picture of the CMOS-sensor.



■ Fig. 31.14

Comparison of the dynamic behavior of a CCD Sensor (*left*) and a nonlinear CMOS Sensor (*right*) (Seger et al. 2000)

Parking and maneuver assistance systems



■ Fig. 31.15

Camera structures and camera system structures (Bosch 2010)

3.4 Structure of Video Cameras and of a Vision System

There are different structures of video cameras, see Fig. 31.15. The camera consists of a lens, an imager, and interfaces. The video data of the first cameras had a standard NTSC-format. Meanwhile, digital cameras with a LVDS interface are mostly used. With the optional LIN/CAN interface, the camera parameters can be controlled by an external CPU.

Smart cameras have an internal Graphic Processing Unit (GPU) and are able to extract features from the picture and thus provide additional information to the driver. Due to the small available space and the thermal conditions, these smart cameras are limited in terms of functionality.

For higher functional demand smart systems are used. They are a combination of a digital camera with an external GPU. The two components can be connected via LVDS. For an around-view system which will be described below, typically four cameras are used.

4 Parking and Maneuvering Assistance Systems

Parking is often difficult for drivers. First of all, an appropriate parking space must be found to avoid unnecessary parking trials. Secondly, the vehicle must be parked, sometimes in unknown environment and without disturbing the surrounding traffic and, sometimes, under observation of other humans. Parking systems can help to find easier an appropriate parking space and to park the vehicle easily, quickly, and safely (Katzwinkel et al. 2009; Kessler and Mangin 2007). As many drivers have problems to see obstacles in the vicinity of the car due to the aerodynamic car body or coverage by the pillars of the vehicle, first solutions to help the driver have been antenna-like bars as early as in the 1990s.

4.1 Requirements to Parking Assistance Systems

A parking system must meet the human requirements with regards to functionality and intelligibility which leads to different requirements to the sensors and the algorithms. The system shall be easy to use and suitable in every situation with the degree of assistance expected by the driver. The human machine interface (HMI) should be designed in such a way that the system behavior can be easily understood by the driver, and it should work in all-days situations. At systems measuring the parking space, the passing speed at the parking space should not be too slow (Blumenstock 2007).

The sensor should fulfill the following requirements:

- Robustness against environmental influences such as rain, dirt, and snow
- High resolution and accuracy of measured distances and parking spaces
- Low cost
- Small installation volume, low power dissipation, and low weight.

The requirements to a system increase with the degree of automation and, thus, the complexity of the system. For systems which recommend the driver a certain parking trajectory or park semiautomatically, the following features shall be considered:

- Robustness against ghost parking spaces (intersections, estate entries)
- Quick system reaction

- The trajectory recommended by the system should be very similar to a trajectory chosen by the human to give the system a high acceptance
- The recommended trajectory must be free of obstacles, and the driver must be warned if there is an object
- The vehicle must be in an adequate final parking position with respect to angle and distance from curb and from objects in the vehicles surrounding
- Short parking time
- Well-designed HMI, easy to operate (Katzwinkel et al. 2009).

The requirements to autonomous vehicles with regards to sensor performance, system, software and hardware performance, reliability, and redundancy are extremely high. Today, the Vienna agreement postulates that a driver must have under all conditions the full control over his vehicle and, thus, it is not allowed to operate a vehicle with a remote control on public roads. Currently EU officials are discussing a change of the Vienna World Agreement.

4.2 Categories of Parking and Maneuvering Assistance Systems

The first product on the market was a reversing aid with ultrasonic sensors and a warning element signaling to the driver the distance to the closest object behind the car with an acoustical and optical warning. This system is the basis for the following generations of ultrasonic-based systems, and the system description will therefore start with warning systems.

In the following paragraphs, the technical realizations of different parking systems are described. They are based on the described sensor technologies and a combination of them. They can be categorized as follows:

Passive systems: They warn or inform the driver without interacting with the cars actuators. They can be subdivided into:

- Warning parking assistance: Ultrasonic-based systems warn the driver from objects behind the car or in front of it and, in the future also, on the sides of the vehicle
- Informing parking assistance: Based on ultrasonic technology, parking-space-measurement systems inform the driver about the length of a parking space and whether it is long enough to park the vehicle or not. If this information is available, a park-steering information system can help the driver to enter best into the parking space
- Simple video-based systems show the driver a picture of a rear-view camera on a graphic screen in the center console area. More sophisticated video-based systems show auxiliary lines within the video picture on the screen giving driving recommendations how to steer best the vehicle into the parking space. Meanwhile, around viewing systems with four cameras are on the market.

Active systems: They inform the driver about possible maneuvers and interact with the vehicles actuators. They can be subdivided into:

- Semiautomatic parking assistance: After measuring the length and depth of a parking space, an optimum trajectory is calculated and the system steers the vehicle into the parking space. The driver has still to care for the longitudinal control of the vehicle
- Fully automatic parking assistance: This type of system releases the driver from both vehicle guiding components, the lateral guidance and the longitudinal guidance of the vehicle by interaction with steering, brakes, and accelerator (Bosch 2009)
- Autonomous parking assistance: These assistance systems take the full control over the vehicle with the driver outside the car. They are in an advanced development state. Some prototype vehicles have been presented (Oertel 2006).

4.2.1 Passive Systems

Passive Systems don't interact with the vehicle; they inform and/or warn the driver from hazardous situations.

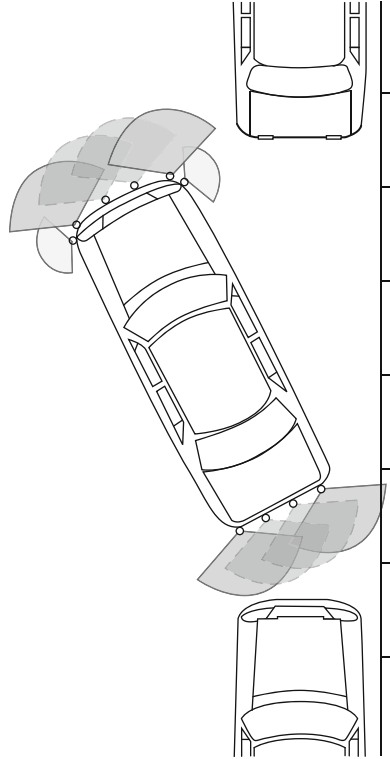
Warning Parking Assistance

Ultrasonic Parking Aid The most common parking aid system is the ultrasonic-based parking aid system. Ultrasonic sensors are well suited for this function and lead to a reasonable cost of the system. The system provides drivers with effective support when parking. They monitor an area of approximately 25–200 cm behind and/or in front of the vehicle, see ➤ Fig. 31.16. Obstacles are detected and brought to the driver's attention by optical and/or acoustic means.

The system comprises the following components: Ultrasonic Sensors, ECU, and appropriate warning elements. Vehicles with rear-end protection normally have only four ultrasonic sensors in the rear bumper. Additional front-end protection is provided by further four to six ultrasonic sensors in the front bumper. High-end vehicles use six sensors on both, vehicle front and vehicle rear.

The detection performance of a sensor or a sensor system can be best evaluated with a Field-of-View (FoV) measurement arrangement. Typically, a tube with a diameter of 7.5 cm is used as a reference object. This is the so-called MALSO-standard for the design of parking systems in passenger cars (ISO 2004). ➤ Figure 31.17 shows the measurement plot of the detection area of a four-sensor arrangement in bird's-eye view.

The system is automatically activated when the reverse gear is engaged or, for systems with additional front protection, when the speed falls below a threshold of approximately 15 km/h or by driver activation via a button. During operation, the self-test function ensures continuous monitoring of all system components. ➤ Figure 31.18 shows the integration of ultrasonic sensor in a bumper of a vehicle.



■ Fig. 31.16

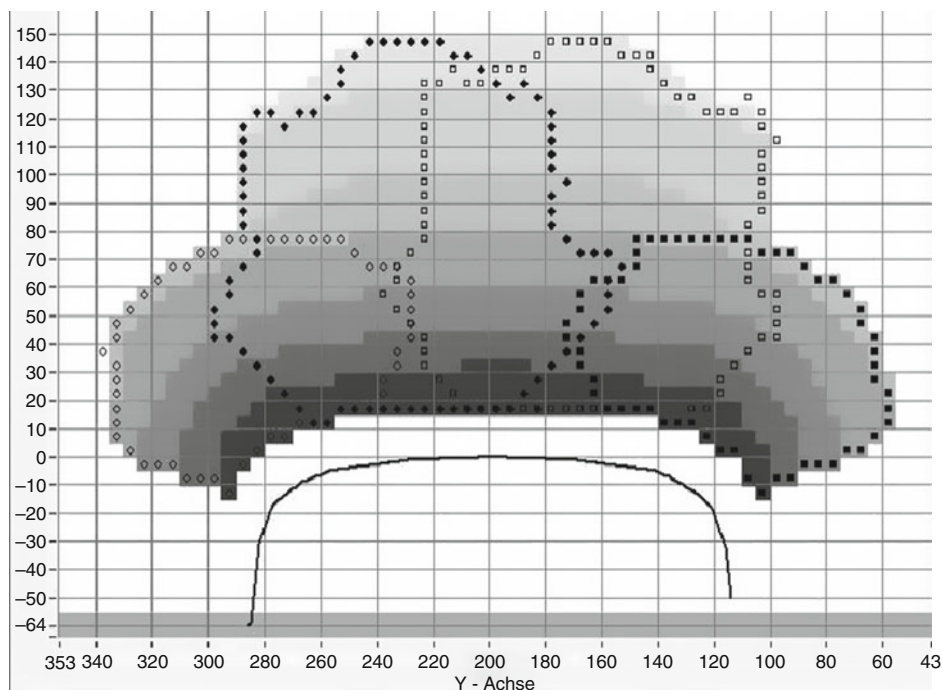
Monitoring range of parking systems with all-around monitoring (Bosch 2011)

Electronic Control Unit The ECU contains a voltage stabilizer for the sensors, an integrated microprocessor (μC), and all interface circuits needed to adapt the different input and output signals, see ● Fig. 31.19.

The software assumes the following functions:

- Activating the sensors and receiving the echo
- Evaluating the propagation time and calculating the obstacle distance by means of trilateration
- Activating the warning elements via an appropriate interface like CAN
- Evaluating the input signals from the vehicle
- Monitoring the system components, including fault storage
- Providing the diagnostic functions.

The latest generations of ultrasonic sensors can be adapted to the specific integration situation of the sensors in the bumper of the vehicle. With this measure, the noise from the road level (clutter) is suppressed even in cases where the ultrasonic sensor cannot be mounted with its membrane vertically to the road surface.



■ Fig. 31.17

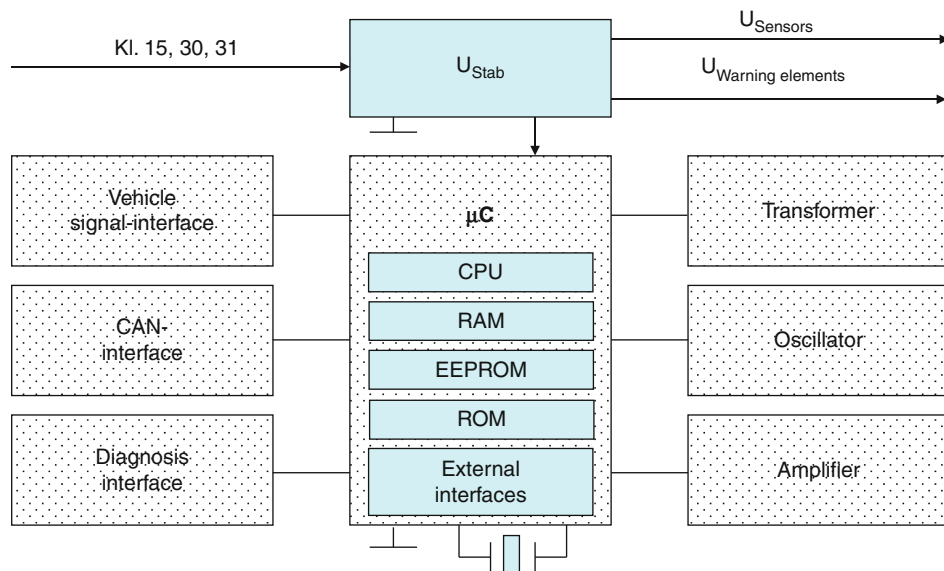
Field-of-view detection field four-sensor arrangement (Noll and Rapps 2009)



■ Fig. 31.18

Integration of ultrasonic sensors into the vehicles bumper (Photo: Bosch)

Warning Elements The warning elements display the distance from an obstacle. Their design is specific to the vehicle, and they usually provide for a combination of acoustic signal and optical display. Both LEDs and LCDs are currently used for optical displays.



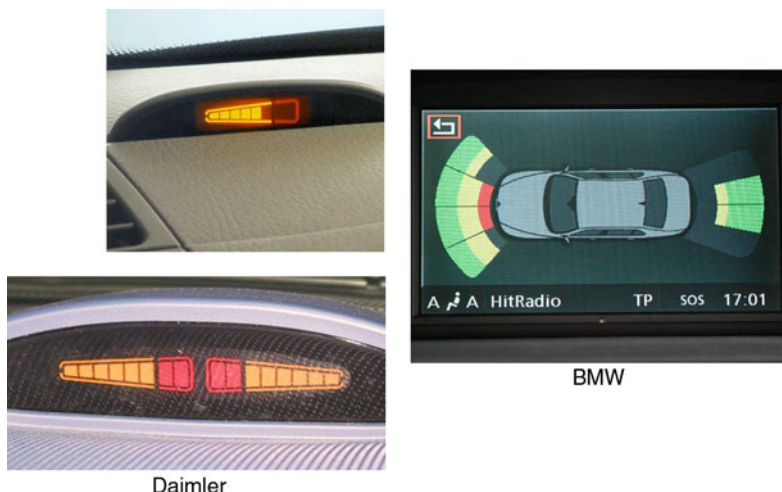
■ Fig. 31.19
Block diagram of ECU of an ultrasonic parking aid (Bosch 2011)

■ Table 31.1
Optical and acoustical output of an ultrasonic parking aid for different measuring ranges (Bosch 2011)

Range	Distance d	Visual indicator LED	Acoustic indicator
I	<1.5 m	Green	Beeping sound
II	<1.0 m	Green + yellow	Beeping sound
III	<0.5 m	Green + yellow + red	Continuous sound
IV	<0.3 m	All LEDs flashing	Continuous tone

With a simple warning element commonly used for the aftermarket segment, the indication of the distance from the obstacle is divided into four main ranges, listed in [Table 31.1](#) see Bosch (2011).

► [Figure 31.20](#) shows two examples of HMI solutions, Mercedes S-Class (left) on the dashboard (upper photograph) and above the rear window, and BMW 7 and 5 series integrated into the center console display. In the latter, obstacles are shown in a bird's-view aspect as green, yellow or red sections within the monitored area in front and in the rear of the vehicle.



■ Fig. 31.20
Examples of warning elements (Photo: Bosch)

Informing Parking Assistance

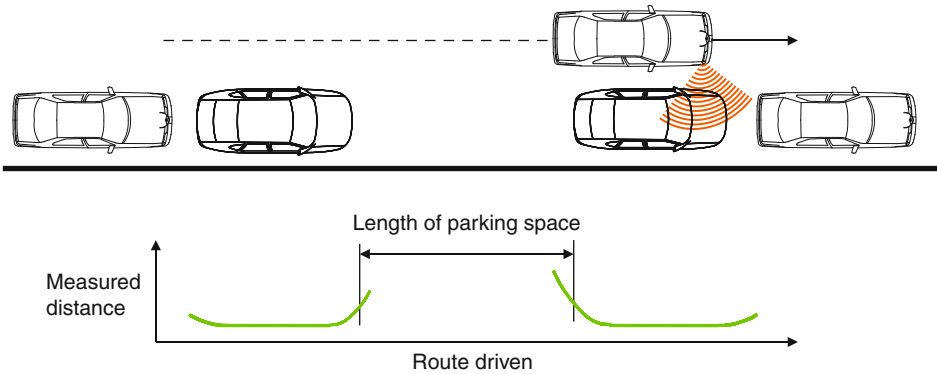
Today, there are five types of informing parking aids:

- The parking-space measurement system
- The park steering information system
- The rear-view camera (picture only)
- The rear-view camera with auxiliary lines
- The video-based around-view system.

Parking-Space Measurement Besides the application of a simple parking aid, the ultrasonic sensors can also be used for the measurement of the size of longitudinal parking spaces. While passing the potential parking space, the length of the parking space is measured. Figure 31.21 shows the principle.

Depending on the length of the parking space, the driver gets an indication whether the parking space is too short or long enough to park into the measured space. Here, binary information, such as “Parking space long enough” or “Parking space too short,” is used. Also, difficulty grades such as “easy,” “normal,” or “difficult” are possible. For this function, the vehicle needs one additional sensor on the front of each side of the vehicle for parking-space measurement and for curb detection.

Automatically (at low speed below 40 km/h) or after the activation of the system by the driver, the lateral sensors scan the parking space while passing it with moderate speed. By using the signals from the wheel speed sensors, the system calculates the length of the parking space and indicates to the driver whether it is long enough to park the vehicle or not.



■ Fig. 31.21
Principle of parking-space measurement (Bosch 2011)

For this use case, sensors with a longer detection range of approximately 4.5 m must be used to allow the system to detect relevant obstacles within the area of the parking space and for curb detection. Should there be an obstacle in the parking space, the system will give the driver appropriate information.

An important factor for the acceptance of such systems is the accuracy of the measurement of the parking space as well as a maximum velocity of up to 30–40 km/h while passing the parking space. “Ghost” spaces such as intersections or estate entries should not be considered as parking spaces.

Parking-space measurement is the basis for the functions “Park steering information” system and “Semiautomatic Parking” described later.

Park Steering Information For vehicles with conventional hydraulic power steering, the Park Steering Information system gives the driver concrete recommendations how to steer the vehicle optimally into the parking space. This requires the following steps:

- Measurement of the parking space
- Calculation of the trajectory
- Continuous calculation of the actual position of the vehicle.

For the calculation of the trajectory, it is advantageous to separate between longitudinal and lateral position. This means that the calculated trajectory consists of strait and curved segments. They can be realized by driving with constant steering-wheel angle and by steering at standstill of the vehicle (Katzwinkel et al. 2009).

For the creation of driving hints to the driver, it is necessary to know the position of the own vehicle relative to the parking space and on the trajectory as precisely as possible. In practice, the localization of the vehicle is made with reference to internal parameters of the vehicle (internal method). This so-called Odometry uses the signals from the vehicles wheels. For the Odometry, the circumference of the wheel is of fundamental important besides the direction of movement of the vehicle because the distance driven is calculated by using this parameter.

The following aspects can influence the circumference of the wheel and must be compensated by appropriate means:

- Production tolerance of the tire
- Wear-off of the tire
- Summer-tires/winter-tires
- Tolerances of the tires.

For a better quality of the position calculation, additional internal parameters can be considered. The steering-wheel angle, the yaw-rate, and acceleration values are usually available on the CAN-bus. These values can be fused with the wheel movement parameters via extended Kalman-filtering.

The following information should be given to the driver during a parking process with steering information:

- Recommended steering-wheel angle or the difference between recommended and actual steering-wheel angle
- Driving direction
- Stopping points for steering activation
- End of the parking process.

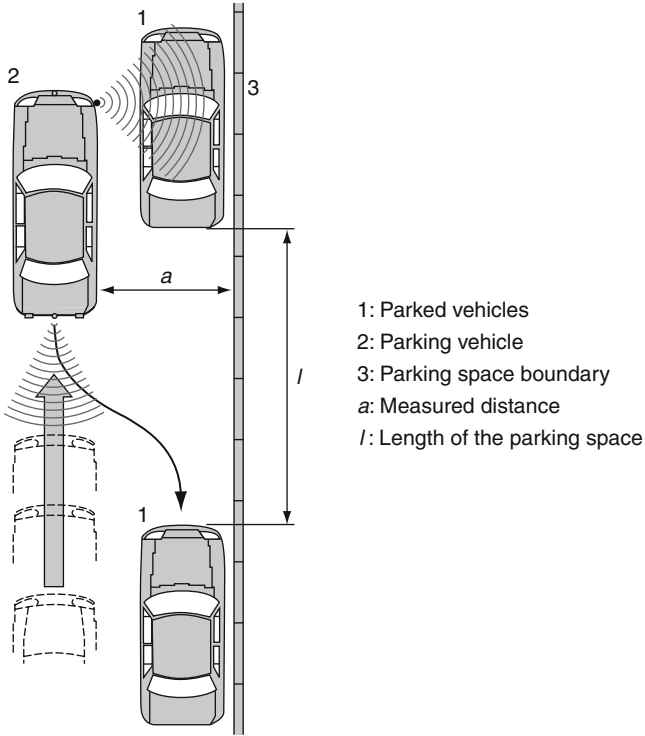
In contradiction to a camera system, the requirements to the display are rather low. A (monochrome) display being available meanwhile in most vehicles is sufficient. To help the driver finding the right steering-wheel angle by giving him information about the recommended and the actual steering-wheel angle, he must step-wise control the steering-wheel angle depending on the current situation. Experiments have shown that this method is better accepted than displaying the difference alone (Sander and McCormick 1987).

After measuring the parking-space length as described above, the ECU calculates the optimum trajectory to park the vehicle from the measured geometry of the surrounding. During entering rearward into the parking space, the system suggests to the driver the optimum amount of steering lock required in order to complete the parking maneuver with the least effort, see ► Fig. 31.22. During the parking process, the trajectory is permanently recalculated and is shown on the display in the dashboard.

During the whole process, the driver has to steer and to care for the longitudinal control of the vehicle, but he is guided into a perfect parking position by clear instructions from the assistant as to steering-wheel position and ideal points at which to stop or to reverse the steering.

Rear-View Camera (Picture Only) Video systems were first introduced in Japan in the 1990s as maneuvering aids for campers, later as reversing cameras for passenger cars showing the picture of a camera with wide angle optics on a display.

The main negative feature of this kind of picture presentation of the vehicle's background is the fact that the extreme wide-angle lens used for this application leads to an extreme picture distortion, see ► Fig. 31.23. This makes it often difficult for the driver to estimate the real distance between vehicle and obstacles. In the vicinity of the vehicle,



■ Fig. 31.22
 Principle of parking steering information (Bosch 2011)



■ Fig. 31.23
 Uncorrected picture of a rear-view camera (Photo: Bosch)



■ Fig. 31.24

Mounting of a rear-view camera at the trunk (Photo: Bosch, from IAA 2009)

obstacles approach much quicker than at longer distances and the risk of a collision is rather high without experience with the system.

The best place for the vehicle integration of the camera with respect to design and environmental influence (mud, snow) is the basin for the grip of the trunk cover or the rear door, see ► Fig. 31.24.

In the following years, new generations of video-based parking systems have been developed.

Rear-View Camera with Auxiliary Lines While the above mentioned informing parking systems give implicit recommendations for driver actions, they are limited by the boundaries of the parking maneuver and do not take into consideration the process of entering the parking space. More information can be provided by a rear-view camera system showing additional auxiliary lines overlaid onto the video picture. To cover a wide detection range behind the vehicle, wide-angle lenses are required. They lead to a significant picture distortion which must be corrected by a picture processing unit to adapt the picture to the visual perception of the human, see ● Fig. 31.25.

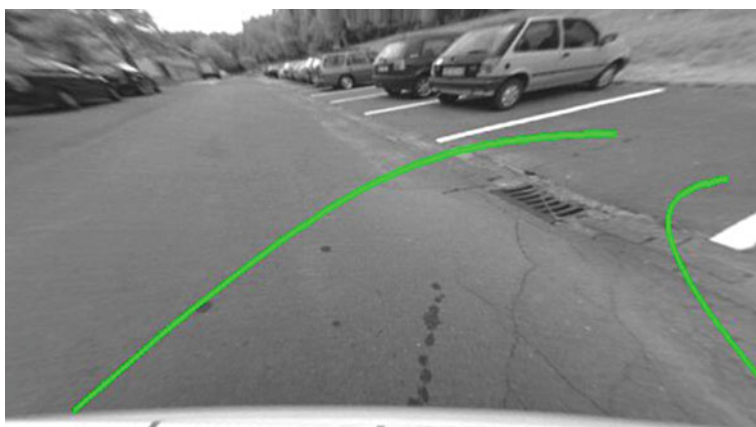
Using systems with additional auxiliary lines, showing the calculated trajectory depending on the actual steering-wheel angle, the driver can be assisted while parking into longitudinal and lateral parking spaces, see ► Fig. 31.26 for a lateral parking space.

More sophisticated systems show additionally lines helping the driver to estimate the right moment for turning the steering wheel to get best into the parking space. In ● Fig. 31.27, an example is shown. When the blue line touches the parking-space marking, the driver must turn the steering wheel completely versus the parking space, and the vehicle follows the red line into the space. In case of longitudinal parking, the driver must turn the steering wheel when the auxiliary line touches the curb.

Around-View System Vehicles with front- and rear-cameras are already available on the market. In addition to this, Japanese legislation postulates the vision of the region besides the codriver in SUV's and trucks making more and more use of cameras instead of



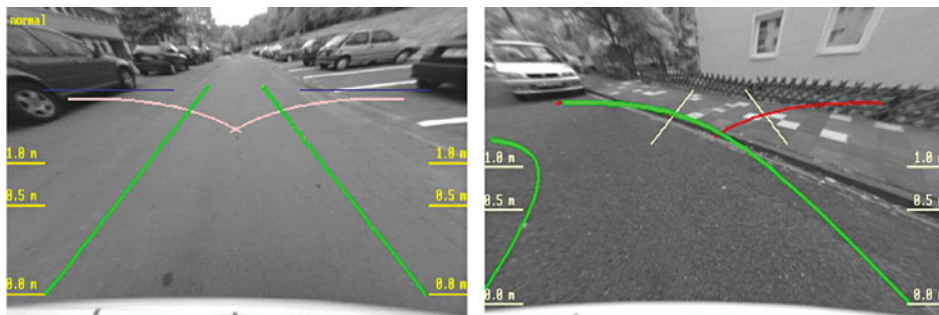
■ Fig. 31.25
Corrected picture of a rear-view camera (Photo: Bosch)



■ Fig. 31.26
Static and dynamic driving corridor prediction for parking into a lateral parking space
(Picture: Bosch Research)

additional mirrors. The cameras can be favorably integrated into the outer rear-view mirrors. Even systems with four cameras are meanwhile on the Japanese and European market.

In the BMW 7-series, a system with four cameras has first been launched on the market. The system provides a passive around-view around the vehicle with different views which can be chosen by the driver. This “virtual” view allows calculating the chosen viewing direction from a fusion of the signals of the four cameras. Distortions caused by



■ Fig. 31.27

Display of auxiliary lines for the definition of the diversion point with longitudinal parking spaces (*left*) and longitudinal parking along the curb (*right*) (Pictures: Bosch Research)



■ Fig. 31.28

Around-View system (BMW Top View) (Photo: Bosch from IAA 2009)

the known fish-eye effect can be corrected in an ECU. The error-free correction into arbitrary viewing directions is only possible under the precondition that the area, on which the objects are positioned, is known. For a bird's-eye view, the road surface is taken as reference. ► [Figure 31.28](#) (BMW 2010) shows an example for this case. Four video cameras with a viewing angle of 190° are arranged on the vehicle (*left*). The right picture shows an example of the viewing direction “bird's-eye view” in a trailer coupling maneuver.

The requirements to the ECU-power are high. The video-data of the four cameras must be composed in real time, and a picture must be generated in a so-called stitching process. Future generations can use the feature of object recognition and classification in the vehicle's surrounding (Gotzig et al. 2009).

The camera-based system can favorably be combined with an ultrasonic system with the ability to measure the distance to objects. This allows the detection of objects which may be poorly visible or obstructed for the camera system.

It is obvious to use a graphic screen as HMI in combination with acoustical warning from the ultrasonic system (Bosch 2006).

4.2.2 Active Systems

Active Systems interact with the vehicle's actuators.

Semiautomatic Parking Assistance

System for Steering into Parking Spaces The next evolutionary step is already on the market. This assistant uses ultrasound sensors to identify suitable parking spaces and acts on the electric power steering unit. All the steering movements are performed by the assistant, while the driver controls the parking maneuver by accelerating and braking.

Park Steering Information, which provides the driver with information on the best possible parking maneuver, remains an alternative technology for vehicles with hydraulic power steering.

During the parking process, the driver must take his hands off the steering wheel. If the driver touches the steering wheel during the semiautomatic parking process, the system switches off and the driver must take over the control over the vehicle.

For this kind of steering interaction, the ECE-regulation 79 (ECE 2006) gives precise rules (Gotzig et al. 2009). The key issue is that the driver must be able to take over the control of the vehicle at any time. As soon as the semiautomatic system is ready, this information must be given to the driver, and the automatic steering control must be switched off automatically if the vehicle speed exceeds the speed of 10 km/h by more than 20% or if the signals needed for the calculation of the action are not any more received by the ECU. If the control stops, the driver must be warned by a clear optical, acoustical, or haptical signal.

The requirements to the HMI are comparable to the requirements for the informing parking systems. ● [Figure 31.29](#) shows an example which is in accordance to the ECE-regulation.

Most semiautomatic parking systems are based on data from ultrasonic sensors and/or from cameras. They show different performance depending on illumination or whether conditions. Compared to ultrasonic systems, weather sensitivity is higher with camera systems due to lens pollution or darkness.

► [Figure 31.30](#) shows the system architecture of a vehicle with a semiautomatic parking system (PSC = Park Steering Control). The ECU of the PSC system (PSC-ECU) is connected with the ECU of the Electric Power Steering (EPS) via the CAN-Bus.

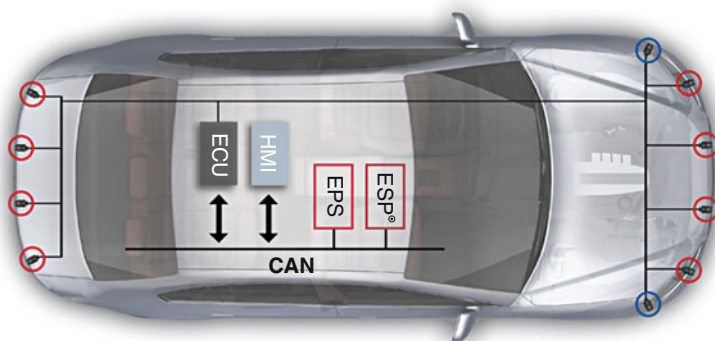
The relevant components are:

- ECU of the parking system
- Switch to activate the system
- Wheel sensors
- Steering-wheel sensor
- Sensors for longitudinal and lateral acceleration
- Lateral ultrasonic sensors for parking-space measurement
- Ultrasonic sensors at the front and the rear of the vehicle for distance measurement to objects



■ Fig. 31.29

HMI of a semiautonomous Parking system (Volkswagen) Photo: Bosch



■ Fig. 31.30

System architecture of a park steering control vehicle (Bosch 2010)

- Turning light switch for the selection of the side of the road to which the vehicle shall be parked
- HMI for interacting with the driver and for distance information of PDC
- Electromechanical power steering
- ECU of the brakes for speed information.

First semiautomatic parking systems are on the market. Meanwhile also those parking spaces can be used needing multiple moves for parking. If the own car comes very close to the vehicle in the rear, the system gives the driver a stop signal. The driver must stop and

activate the forward gear. The system makes a full turn of the steering wheel and corrects the position of the vehicle within the parking space. This function meanwhile works very reliably for longitudinal parking spaces and is currently extended also to parking spaces located at a right angle to the driving direction (Katzwinkel et al. 2009).

New Assistant for Driving out of Parking Spaces The system for parking into parking spaces can also be used for parking out. It helps drivers get into the right starting position first and will then carry out the steering maneuvers required to pull out of a parking space safely. The driver's role is to keep his eye on traffic, and step on the gas pedal and the brakes at the right moment. To turn the function off, all the driver needs to do is grab the steering wheel.

Fully Automatic Parking Assistance

The fully automatic parking Assist is under development (Bosch 2009). The system has the complete functionality of the semiautomatic parking assistance system, but it cares additionally for the longitudinal control. While entering a parking space the system steers, accelerates, and brakes the own vehicle depending on the position of other cars and obstacles. All the driver has to do is switching on the reverse gear at a detected parking space and to put his hands on the lap – the rest is done by the system. However, the driver is still responsible for the parking maneuver and therefore has to monitor the situation to take over if required.

Autonomous Parking Systems

During the European PROMETHEUS-project (*PROgramMme for a European Traffic with Highest Efficiency and Unlimited Safety*) already in 1990, fully autonomous parking vehicles have been presented by different research divisions of large automotive companies and suppliers. Today, the infrastructure of modern vehicles is designed in such a way that all components are networked with each other by the CAN-bus, can exchange data, and be influenced via the bus. Nevertheless, there is still no series application. In Oertel (2006), autonomous parking without the driver sitting in the vehicle has been shown recently. The driver starts the parking process from outside the vehicle by pressing a key on the cars key, and the vehicle slides slowly into the garage but only if the key is kept activated by the driver. The system is based on ultrasonic sensors and cameras. Such a system could be an interesting feature in garages with narrow parking spaces and on private grounds. The legal requirements, such as product liability, to use such a system on public roads have to be clarified. Today, autonomous systems without the driver having his full control over the vehicle are forbidden by the Vienna World Agreement.

5 Benefit of Parking Systems

It is known that there are 4,000,000 material damage accidents per year in Germany with 6 Billion EUR damage cost. Unfortunately, accidents with little damage volume are not registered by statistics and, thus, there doesn't exist any statistical material from the

insurance companies about cost savings with these systems. But it can be assumed that the benefit is significant as repairs at bumpers are rather expensive and that the cost of the investment usually pays by far for itself during the life of a vehicle.

The successful story of ultrasonic technology in comparison to radar, infrared, capacitive, or inductive technologies is based on a couple of features, making it superior to other measurement technologies. The main factors are:

- Low production cost
- High robustness against environmental parameters
- Good detection quality almost independent from the kind of obstacle. Relevant materials like metal, plastic, wood, brick walls, or glass are “hard” for sound waves at their surfaces and give almost identical echo signals.

Due to picture presentation, video systems have much higher information content than ultrasonic systems. The advantage of presenting a picture to the driver results in a quick object classification and a much easier decision if there is a relevant object or not.

The systems described above lead to a significant increase of driving comfort. Especially inexperienced drivers estimate the assistance given by the systems. Furthermore, vehicles are removed quicker from the flowing traffic and, thus, the risk for an accident decreases.

6 Functional Limits of Parking Assistance Systems

Ultrasonic technology has some restrictions in functionality:

- Sound absorbing materials (e.g., plastic foam) are hardly seen by the system. But, in practice, this drawback plays no role
- For persons wearing absorbing cloths (e.g., woman with a fur-coat), the system has a shorter detection range
- There may occur acoustical interference from other objects in the vicinity of the own vehicle, in particular the noise of compressed air (e.g., truck brakes) and metallic friction noise from track vehicles
- The sensor may be covered with mud or snow under severe weather conditions. The detection range may therefore decrease or detection may be impossible for the system. In such a case, the system detects its malfunction and gives the driver an acoustical and optical warning via the HMI of the vehicle.

Video technology has also restrictions:

- Poor weather conditions like fog or rain may reduce the visibility range of cameras based on Silicon sensor technology significantly
- Camera lenses may be covered with mud or snow at poor weather conditions and must be cleaned from time to time.

Due to their limited performance, Parking systems based on ultrasonic sensors and cameras are therefore defined as comfort systems.

Ultrasonic sensors and video cameras depend on completely different physical principles and are therefore an excellent supplement to each other. Each technology has individual strengths supporting the weaknesses of the other. The camera-based system can therefore favorably be combined with an ultrasonic system with the ability to measure the distance to objects. This allows the detection of objects which may be poorly visible or obstructed for the camera system and the video picture with integrated information for an ultrasonic parking system contains much more information for the driver. This is an important step towards more detection security and functional safety.

Radar sensors have a longer detection range and may be mounted behind the plastic bumper fascia. They may be used for parking aid as well as for safety functions such as collision avoidance or collision mitigation on the basis of a multiuse concept.

7 Legal Aspects of Parking Assistance Systems

There are a couple of legal aspects to be considered during the design and operation of Driver Assistance Systems in general and Parking Assistance Systems in particular.

In order to avoid liability problems, the systems should be classified in:

- Informing/assisting systems (“aid”). The vehicles manual must contain all limitations and operational restrictions the system may have
- Semiautomatic or automatic systems with easily understandable interaction of the system. As uncritical those systems are seen which do not distract the driver (e.g., ESC, Electronic Skid Control)
- Automatic Systems with not easily understandable interactions (e.g., emergency braking while the driver tries to avoid a collision with evasion). They are not allowed according to the Vienna World Agreement.

For all kinds of Driver Assistance Systems, the following rules apply:

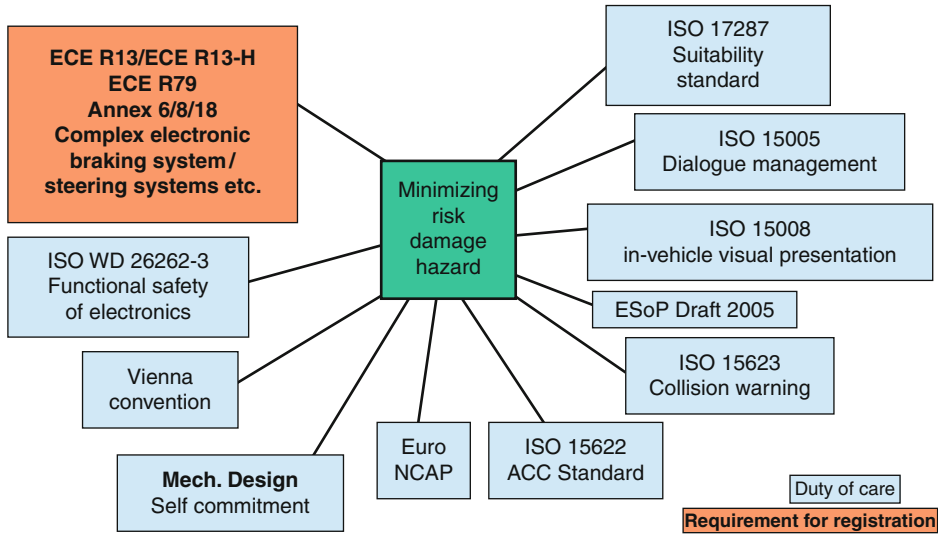
- The system shall only intervene if necessary
- The system may not lead somebody to believe that there is a system safety
- The driver shall not be lulled by the system
- The system shall not distract the driver’s attention.

These are the general rules for the design of the system components regarding functionality and Human-Machine Interface design.

The Vienna World Agreement was signed 1968 under the roof of the United Nations by 63 countries with the goal to improve road safety. The agreement is an obligation of all participating countries to issue common rules for the road traffic and for the admission of vehicles to the road traffic (Lambert et al. 2008).

The articles, regulating the guidance of vehicles on public roads, are Articles 8 and 13. Article 8 says:

- “Every driver shall at all times be able to control his vehicle or to guide his animals.”



■ Fig. 31.31

Current rules and regulations for driving assistance systems (Knoll 2010)

Article 13 says:

- “Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all maneuvers required of him.”

The European Product Liability Law says:

- “A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account, including:
 - The presentation of the product
 - The use to which it could reasonably be expected that the product would be put
 - The time when the product was put into circulation (EC 1985).”

➤ *Figure 31.31* shows the current rules and regulations in conjunction with Advanced Driver Assistance Systems (ADAS) as an overview.

8 Conclusion

Both described technologies, ultrasonic and video, have reached a high development state, but there is still significant potential to develop with the sensor technologies themselves as well as with new functions based on them. First to mention ultrasonic technology.

Piezoceramic ultrasonic sensors for automotive parking systems have been improved significantly since their first series application in 1992 regarding their mechanical,

acoustical and electronic features. Today's sensors are compact and robust, and may be easily and hardly visibly integrated into varnished or chromium bumpers. Their acoustical behavior is optimally adapted to the angular characteristics, and they can be electronically adapted to the individual needs of the customer and the integration needs into the vehicle. Further improvement of the measuring system for advanced parking and maneuvering systems are expected in the future.

With regards to video technology, the following development potential can be defined:

- Quick cost decrease due to a wide commercial application
- Imagers with higher resolution are soon available at reasonable cost
- Due to the availability of powerful microprocessors, new picture processing methods are being developed, especially in the field of pedestrian detection and protection.

Sensor data fusion of ultrasonic and video sensors opens the vision for the realization of more powerful systems. For a fusion of the ultrasonic system with the Vision system, video data are processed and fused with the information of the ultrasonic system. Using this complementary and partly redundant sensor information, such a system can be developed to a thorough parking and maneuvering system providing more extensive information for the driver to better control his vehicle (Denner 2010).

In general, benefit and acceptance of all previously described functions, besides reliable and robust sensors, depend extremely from the data processing and from a good HMI concept. These development blocks are under a permanent improvement process. All these factors are the basis for a successful further market penetration and for the introduction of new and innovative functions for ultrasonic-based and video-based Driver Assistance Systems.

References

- Blumenstock KU (2007) Platz da? Vergleich von fünf Einpark-Assistenten (Comparison of five Parking Assistance Systems) Auto, Motor und Sport, Vol. 13
- BMW (2010) www.bmw.de. Status 10/2010
- Bosch R (2006) Safety, comfort and convenience systems. Wiley, Chichester
- Bosch R (2008) Assistance systems. The Bosch Yellow Jackets, Stuttgart
- Bosch R (2009) On the road to fully automatic parking. Press Release, August 2009
- Bosch R (2011) Kraftfahrzeugtechnisches Handbuch (Bosch Handbook of Automotive Technology), 27th edn. GWV-Fachverlage, Wiesbaden
- Denner V (2010) Vernetzung und Elektrifizierung – Neue Herausforderungen für die Automobiltechnik (Networking and Electrification – New Challenges for Automotive Technologies). In: Proceedings of Fortschritte in der Automobiltechnik (Progress in Automotive Technology), Ludwigsburg
- EC (1985) Directive on Product Liability 85/374/EEC
- ECE (2006) Regulation 79 Rev. 2, 20
- Gotzig H, Mathes J, Barth H (2009) Die naechste Generation Einparksysteme - Entwicklungsstufen, Studien und Trends (The next generation of Parking Assistance systems – development trends, studies and trends), ATZ, Dec 2009
- ISO (2004) ISO17386:2004(E) Maneuvering aids for low speed operation (MALSO)
- Katzwinkel R, Auer R, Brosig S, Rohlf M, Schoening V, Schroven F, Schwittes F, Wuttke U (2009) Einparkassistentz (Parking assistance). In: Winner H et al (eds) Handbuch Fahrerassistenzsysteme (Handbook of driver assistance systems). GWV-Fachverlage, Wiesbaden

- Kessler M, Mangin B (2007) Nutzerorientierte Auslegung von teilautomatisierten Einparkassistenzsystemen (User-oriented Design of semi-automatic Parking Assistance Systems), 4. VDI-Tagung Fahrer im 21. Jahrhundert, Braunschweig
- Knoll PM (2003) Video sensors. In: Marek J et al (eds) Sensors for automotive technology. Wiley-VCH, Weinheim
- Knoll PM (2010) Predictive driver assistance. Lecture at KIT (Karlsruhe Institute of Technology)
- Lambert G, Kirchner A, Hueger P (2008) Parkassistentensysteme – Technologien von heute und morgen (Parking assistance systems – technologies today and tomorrow). VDI-Verlag, Düsseldorf
- Noll M, Rapps P (2009) Ultraschallsensorik (Ultrasonic sensor technology). In: Winner H et al (eds) Handbuch Fahrerassistenzsysteme (Handbook of driver assistance systems. GWV Fachverlage, Wiesbaden
- Oertel K (2006) Garagenparker von BMW (Garage parking device from BMW), Hanser automotive, 5-6/2006. Carl Hanser Verlag, München
- Sander MS, McCormick E (1987) Human factors in engineering and design. McGraw-Hill, New York
- Seeger U, Knoll PM, Stiller C (2000) Sensor vision and collision warning systems. In: Convergence international conference, Detroit
- Waanders JW (1991) Piezoelectric ceramics, properties and applications. Philips Components Marketing Communications, Eindhoven, Netherlands

32 Post-crash Support Systems

Jeffrey S. Augenstein¹ · George T. Bahouth²

¹Miller School of Medicine, University of Miami, Florida, USA

²Impact Research, Inc., Columbia, USA

1	<i>Introduction</i>	866
2	<i>Opportunities for Improved Post-crash Care</i>	869
3	<i>Estimating Crash Injury Risk Following a Crash</i>	872
4	<i>Injury Risk Threshold Selection</i>	874
5	<i>Future Opportunities to Improve Post-crash Care</i>	878
6	<i>Conclusion</i>	879

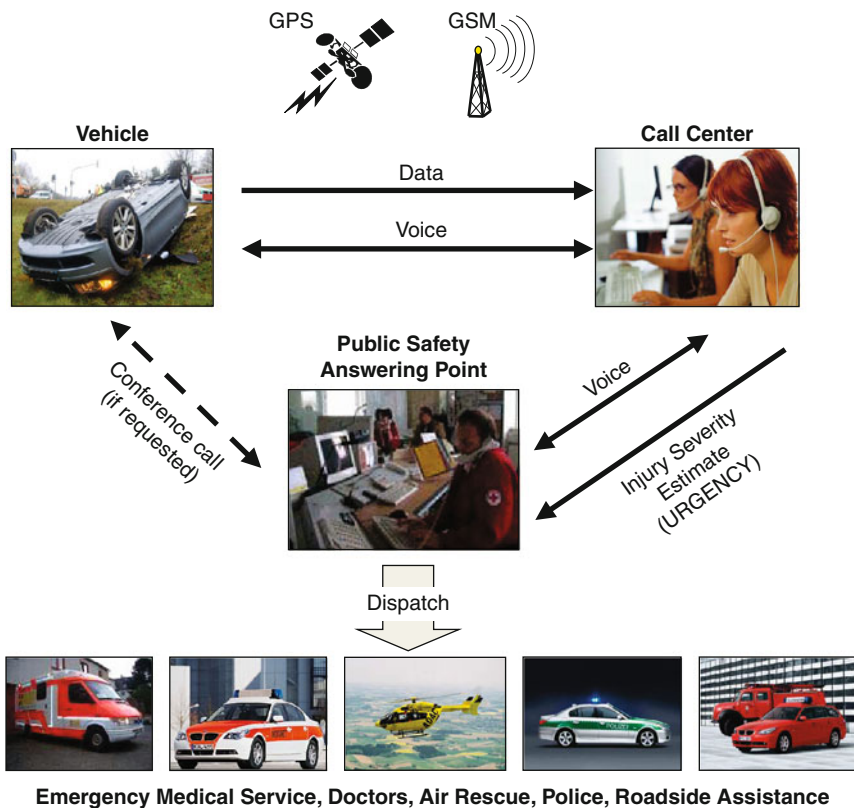
Automatic Crash Notification (ACN) systems employ advanced vehicle sensor data and in-vehicle communications technology to facilitate more efficient and effective post-crash support for drivers and passengers. An immediate call for help coupled with accurate information on location and crash severity could positively impact multiple levels of the transportation and rescue system. First, and most importantly, the rapid notification of rescue services and appropriate treatment of occupants following a serious crash has been proven to be a critical factor in reducing patient morbidity and mortality. These systems have the ability to save lives and directly improve patient care. Second, knowledge of the characteristics of a crash will allow rescue services to dispatch the most appropriate resources, which becomes critical in areas with limited resources or in areas where the crash site is at a significant distance from the nearest treatment facility. Finally, immediate recognition of a crash and use of this information by highway officials may help to dynamically reroute traffic or initiate the clearing of a scene in order to reduce traffic congestion associated with crashes.

1 Introduction

As early as 1995, motor-vehicle manufacturers began equipping cars with basic ACN technology to transmit a mayday signal in the event of a moderate-to-high severity crash. Qualifying crashes are those severe enough to deploy airbag systems or other safety restraint technologies. In the earliest vehicles, the ACN signal contained vehicle identification data, exact GPS coordinates, and vehicle airbag deployment data. In addition, a voice link is established between the car and the telematics service provider (TSP) just moments after the crash so that additional information can be gathered and appropriate support decisions can be made. Some existing ACN systems use an alternative approach where voice communications are directly established with 911 services without an initial call to a TSP first.

In recent years, Enhanced ACN systems (also known as Advanced ACN Systems or eACN) have emerged that transmit location and crash feature data like the basic ACN system but also transmit crash characteristics compiled by in-vehicle sensor systems. These characteristics include well-established factors needed to assess crash severity. At the present time, these factors include the crash energy commonly referred to as ΔV , impact direction, presence of a right front passenger, knowledge of three-point belt usage in front seats, occurrence of a rollover event when sensor data is available, and the recognition of multiple impact crash events. ➤ [Figure 32.1](#) shows the sequence of events that occurs following a crash and the interaction between the principal system components.

As shown, once a crash occurs that is severe enough to deploy airbags or other safety systems, the vehicle will send an automatic signal to the TSP. This signal includes vehicle identifiers, the location of the crash, and key crash attributes. At the same time a voice link is established between the occupants of the crashed car and the call center. Based on the information transmitted automatically by vehicles and the information gathered verbally,

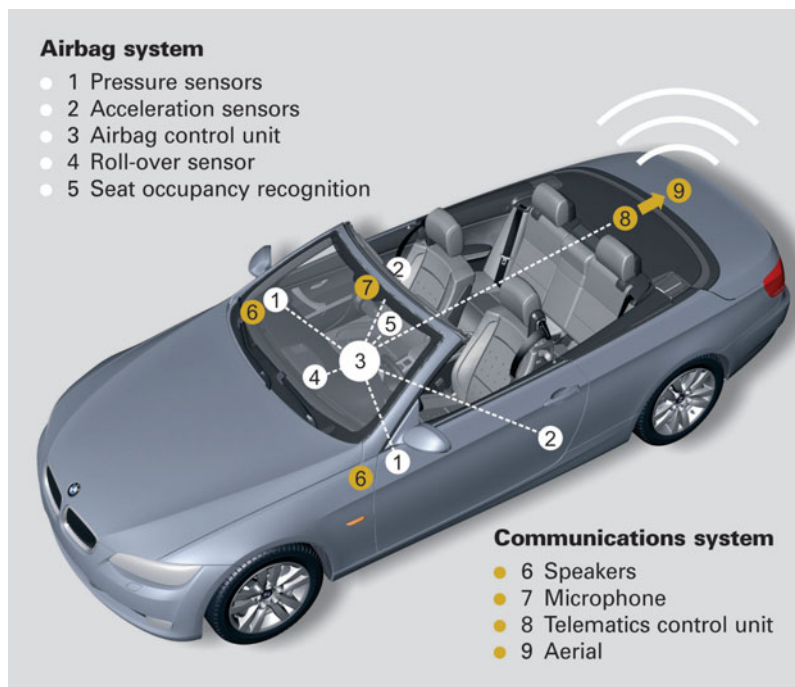


■ Fig. 32.1

Rescue sequence including the enhanced automatic collision notification system

the call center operators may choose to contact the Public Safety Answering Point – PSAP/ 911 center – to relay location and severity information to the call taker for further action. As previously mentioned, some manufacturers currently opt to establish the initial call directly with the PSAP once airbags have deployed and drivers are given an opportunity to opt out of the call if they choose to do so. Based on the information shared and dispatch protocols they use, a subsequent dispatch of police, fire, Emergency Medical Services (EMS), or roadside assistance may occur. If requested, most ACN systems allow a voice link to be established directly between the occupants of the vehicle and the PSAP.

The principal components used by the ACN system are listed in ● Fig. 32.2 including accelerometers, pressure sensors, and gyroscopic sensors. If a crash event exceeds the predetermined threshold for transmission of an ACN signal, verbal communication occurs between the TSP and the occupants through a fixed microphone and the vehicle audio system. In any other emergency situations, a manual emergency call can be triggered by the customer via a dedicated SOS button. Use of the button does not interfere with an automatic emergency call.



■ Fig. 32.2

Principal vehicle components used by the ACN system

Although technology solutions vary between manufacturers, most systems use a built-in Telematics Control Unit with integrated GSM technology. These integrated systems do not require users to have a handheld mobile phone not only available but paired with the vehicle in order to use the system.

The vehicle position and heading are critical for rescue and are continually calculated and monitored within the in-vehicle system. Even if GPS reception is temporarily unavailable, for example, in tunnels, and natural or manmade obstructions, highly accurate positioning is achieved by dead reckoning and, if a navigation system is present, using map-matching techniques. The system could also maintain a list of recent significant way points, which are included in the eACN data-packet. Recent way points can often assist in accessing the exact vehicle location, for example, in the case of complex road junctions, bridges, or in case of close parallel roads and/or motorways. In the unlikely event that the positioning system or GPS antenna is damaged during the crash, the above techniques help ensure that the in-vehicle system still has access to the last-known positions from only seconds beforehand.

Current eACN in-vehicle systems use GSM SMS technology in the United States and Europe for data exchange. Retry mechanisms are included to improve data transmission reliability, if necessary. If the system detects that a call cannot be connected, or that an active call has been dropped, then automatic retry mechanisms are incorporated.

Alongside the transmission of an eACN data-packet, verbal communication between the TSP and the occupant occurs through a directly connected built-in microphone and speaker system. Once connected, each call taker can verify location and nature of the crash before appropriate rescue decisions are made.

Overall, the goal of this highly reliable and redundant safety system is to identify crashes and occupants most in need of rescue assistance.

2 Opportunities for Improved Post-crash Care

In the event of a crash where serious injuries occur, a rapid call for help followed by appropriate emergency medical services may significantly improve medical outcomes for drivers and passengers. The traditional approach to deploying life-saving services to a crash has been based on communication to the appropriate 911 system by occupants of the crashed vehicle and/or external observers. The advent of cellular telephones enhanced the recognition of crashes. Unfortunately, it is not uncommon for multiple calls to be presented to the 911 system each describing different severities, apparent injuries, and even locations for the same crash. In the case where a crash occurs in a remote area or is unobserved by passing traffic, seriously injured and potentially disabled occupants would most benefit from an automatic call for help.

In the United States in 2009, there were approximately 5.5 million drivers involved in motor-vehicle crashes reported to police (NHTSA 2010). Of those, 1.5 million drivers are injured while 110,000 sustain serious injuries requiring immediate medical attention. An important goal of the ACN system is to discriminate those with severe injuries from those without injury automatically based on vehicle telematics data transmitted immediately following a crash. Subsequently, this information could be used by rescue dispatch and rescue personnel on scene to improve care for crash victims.

The rapid identification of critically injured occupants followed by appropriate care has been shown to improve injury outcomes and prevent fatality. A study by Clark and Cushing suggests a 6% fatality reduction is possible (1,647 lives in 1997) if all time delays for notification of EMS were eliminated even if methods for dispatch and treatment remained the same (Clark and Cushing 2002). This reduction in notification time would occur with widespread implementation of eACN/AACN technology in passenger vehicles today.

Three studies conducted by the NHTSA have explored preventable deaths to assess the effectiveness of the current trauma care system (Esposito et al. 1992; Cunningham 1995; Maio et al. 1996). Two of the studies concluded that 28.5% and 27.6% of fatalities occurring in their regions were preventable with improved EMS and treatment. The third study concluded that 17% of fatalities occurring in combined urban and rural areas were also preventable. Delayed treatment and improper management of the injured were cited as the factors that most frequently contributed to the avoidable death. The majority of the preventable deaths occurred after arrival at a hospital. This study suggests that opportunities exist for preventing trauma deaths not only by reducing the time from

crash to hospital, but also by aiding in the recognition of the nature of the most serious injuries. A more recent study conducted in Victoria Australia identified that most crash victims (85.4%) that die before reaching a hospital do so because of a major injury that is considered unsurvivable (Ryan 2004). However, 14.6% were identified as possibly preventable. Further, this study did not evaluate the effect of improved care on occupants who arrived at hospitals and subsequently died.

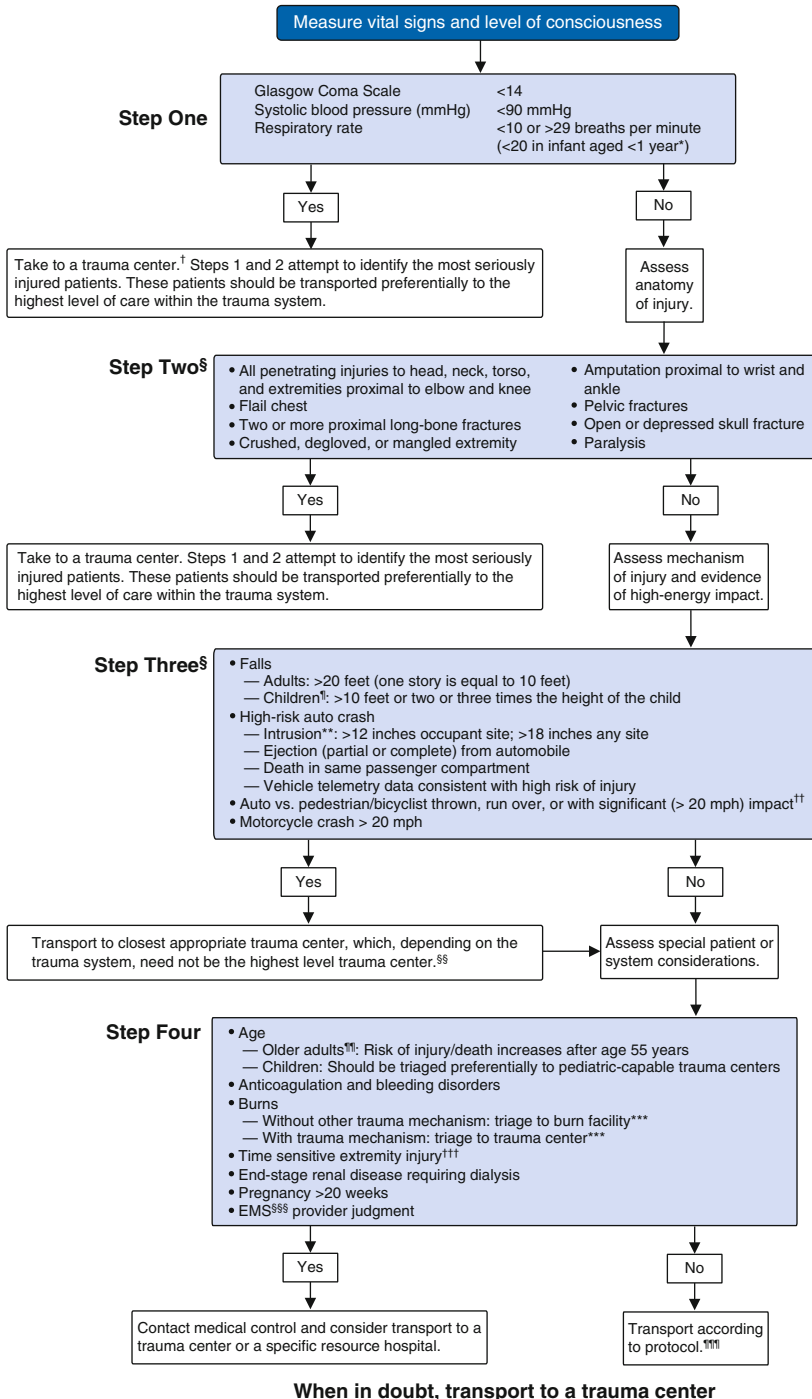
A recent evaluation considering the effect of trauma center care on mortality of patients arriving at hospitals with one or more Abbreviated Injury Scale Level 3 injury (AIS 3) underscores the importance of treatment in the most appropriate medical facility. Overall, the findings of this study suggest that the risk of death is 25% lower when care is provided in a trauma center compared to a non-trauma center (MacKenzie et al. 2006). A second study highlighted that patients taken to a Level I Trauma Center versus Level II hospitals not only had better survival rates but improved functional outcomes (Cudnik et al. 2008).

Currently, there is a problem with overtriage of a large proportion of trauma patients. This is perhaps most accentuated with victims of motor-vehicle-related crashes. The pre-hospital criteria used to determine who needs to be triaged to a trauma center after a motor-vehicle crash are poor predictors of injury. As many as 50% of blunt trauma patients initially thought to be seriously injured and transported to a trauma center are not admitted or are discharged from the hospital within 24 h (Norwood et al. 2002; Kohn et al. 2004). Better triage of these patients will lead to better resource utilization and cost savings.

Field triage criteria have been established by the American College of Surgeons (ACS) and include physiologic, anatomic, and mechanism of injury criteria to justify transport directly to a trauma center. Since no criteria will capture all severely injured patients, most trauma systems rely on “paramedic judgment” or “high suspicion of injury.” Paramedic judgment is used to assess the scene and suspect serious injury based on the crash mechanism and crash characteristics. Unfortunately, this has led to a large overtriage rate, worsened by an increasingly novice workforce of paramedics. This means even in the absence of defined trauma triage criteria, the first responder can use his/her judgment to upgrade the patient to a trauma alert. This is also reflected in the fact that an overtriage rate of 25–50% is considered acceptable by the ACS in order to ensure that all severely injured trauma patients are seen at a trauma center (ACS 2006).

In 2009, the US Centers for Disease Control worked jointly with the National Highway Traffic Safety Administration (NHTSA) to develop a revised Guidelines for Field Triage of Injured Patients. The result was a decision scheme divided into four steps examining (1) Physiologic criteria, (2) Anatomic criteria, (3) Mechanism of injury, and (4) Special patient or system considerations (see ● Fig. 32.3). At each step, the decision scheme presents two transport alternatives determined if the patient’s condition is serious enough to require transport to a certain level of trauma care.

One criteria of particular importance for post-crash safety is the “High-risk” auto crash category indicating that the “Vehicle telemetry data consistent with high risk of injury” criterion has been met. While this criterion is not overly specific, it allows for the



■ Fig. 32.3

Field triage decision scheme – Proposed by US centers for disease control, 2006

use of eACN data elements to assess the risk of severe injury and application of this risk data to assist rescue in their trauma triage decision-making process.

Below, an advanced, instantaneous algorithm known as URGENCY is described. Its purpose is to predict severe crash injury risk and to assist in dispatch and trauma triage decision making and can aid the pre-hospital provider in their efforts to properly triage motor-vehicle crash victims to a hospital or trauma center and to assist in-hospital staff during patient evaluation and treatment. It is meant to be used as an adjunct to the already defined trauma triage criteria and to provide some objectivity to the use of “paramedic judgment” as a criterion. The incorporation of an algorithm based on eACN/AACN data may reduce the overtriage of patients when other criteria for trauma team activation are either not met or unavailable.

3 Estimating Crash Injury Risk Following a Crash

In order to identify crash events where a severe injury is likely, key characteristics describing the crash configuration, crash energy, and occupant factors are important. Key crash attributes that best discriminate seriously injured occupants include the direction of impact for each event (frontal, nearside, farside, rear impact, or rollover), the impact severity based on deltaV for each impact, the use or nonuse of safety belts in front seats, and the number of impact events that occurred (Farmer et al. 1997), (Cummings and Rivara 2004), (Bedard et al. 2002) (Jones and Champion 1989). Secondary variables that could be collected through voice communications include occupant age and gender for drivers and passengers.

The URGENCY Algorithm consists of a series of logistic regression models which relate the risk of high severity injury to a series of independent variables describing the crash event. Although the definition of a severely injured occupant can vary between systems, the definition includes occupants who sustained one or more injury with an Abbreviated Injury Severity (AIS) Score of 3 or higher during a crash due to trauma (includes AIS 3, AIS4, AIS 5, and fatally injured). This group is referred to as MAIS3+ injured.

Data from the National Automotive Sampling System Crashworthiness Data System (NASS CDS) was used initially during model development to relate crash characteristics to the risk of serious injury for target occupants. NASS CDS is collected by the NHTSA and contains a sample of 4,500–5,000 crash cases annually (15). Each case involves at least one motor vehicle in transport on a public roadway where one or more vehicles were towed from the scene. Data elements recorded in NASS CDS cases are collected by professional crash investigators based on an in-depth inspection of the vehicle interior, exterior, and the crash scene. Supplemental information is gathered from police reports, occupant interviews, and hospital records.

Each NASS CDS case is assigned a weighting factor to reflect its probability of sampling. Case weight adjustments were made to reduce the impact of outlier weights on injury rates. This process is an important step in order to reduce the variability between cases with each stratum. When weighted before and after adjustment, the sample represents the nationwide incidence of tow-away crashes and resulting injuries.

The statistical software SAS Version 9.1 was used for data handling and Stata Version was used to compute parameter estimates and standard errors due. Stata was necessary to accommodate the stratified sample of cases within NASS/CDS.

This study addresses passenger vehicle front seat occupants over the age of 12 who are involved in planar only crashes from 2000 to 2007. Model year 1998 and later vehicles only were used during model development and evaluation. The weighted and unweighted crash populations used during model development are shown in [Table 32.1](#) below.

Each model was subsequently evaluated using a population of NASS CDS cases randomly selected from the complete population of 2000–2007 crash data (see [Table 32.1](#) for evaluation dataset population sizes). These cases are independent of those used to train the model and were analyzed to determine the predictive value of the models. [Table 32.2](#) shows the overall ability of the models to identify or capture the MAIS3+ injured within the evaluation population (i.e., model sensitivity). Further, the table presents model specificity which indicates the model's ability to capture the uninjured within the evaluation population as well. These values are presented in [Table 32.2](#) for planar only crashes by crash direction. [Table 32.2](#) shows model parameters by crash direction. These estimates, when applied to each unique case, can be used to estimate the risk of severe injuries for a particular front seat occupant.

This modeling strategy assumes that TSP has instant access to vehicle data and will rapidly assess the risk of MAIS3+ injury with or without verbal response from occupants. To highlight the difference in injury risk based on seat beltedness and principal crash direction, [Figs. 32.4](#) and [32.5](#) are offered. As shown, there is significant variability in risk comparing rear crashes and nearside crashes. Establishing that a nearside collision has occurred required the combined use of the Principal Direction of Force (PDOF) and seat occupancy.

For unbelted occupants, the risk of severe injury varies by crash direction however less so than that of the belted population.

The overall predictive accuracy of the model suggests that 75.9% of injured occupants would be correctly identified using data automatically collected and transmitted by vehicles alone. In other words, an automatic call for help indicating serious injury is likely to be made for three out of four MAIS3+ injured occupants even if their crash

■ **Table 32.1**

Crash populations used for URGENCY algorithm development (NASS CDS 2000–2007)

Population	Unweighted count	Weighted count
Training dataset		
Non-MAIS3+ injured	23,655	13,270,000
MAIS3+ injured	3,474	339,500
Evaluation dataset		
Non-MAIS3+ injured	4,940	3,109,000
MAIS3+ injured	880	92,500

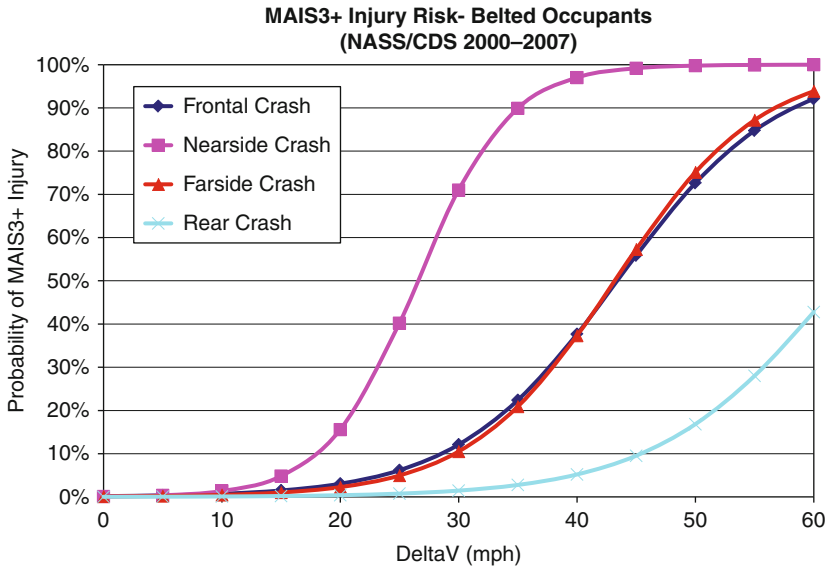
■ Table 32.2
URGENCY parameters by crash direction

Crash mode	Parameter	Estimate	Standard error
Frontal	Intercept	−6.2697	0.149
	DeltaV(MPH)	0.1474	0.00540
	Belt use	0.0311	0.0904
	Age	−0.8055	
	Multiple impact (Y/N)	0.4001	0.0908
Nearside	Intercept	−5.2836	0.0299
	DeltaV(MPH)	0.2102	0.00111
	Belt use	0.0309	0.0201
	Age	−0.726	
	Multiple impact (Y/N)	1.2299	0.0155
Farside	Intercept	−4.7847	0.328
	DeltaV(MPH)	0.1505	0.0128
	Belt use	0.0254	0.205
	Age	−1.5644	
	Multiple impact (Y/N)	0.8037	0.219
Rear	Intercept	−5.6349	1.002
	DeltaV(MPH)	0.1337	0.0338
	Belt use	0.0295	0.684
	Age	−1.6751	
	Multiple impact (Y/N)	0.1974	0.628

was not observed by somebody on scene or if occupants were unable to place a call themselves. These results are shown by crash direction and overall in ► Table 32.3. When URGENCY estimates are used in combination with verbal information gathered by the TSP or 911, occupants in need of medical attention would be rarely missed. A third opportunity to assess injury severity exists before hospital transport once EMS has arrived on scene.

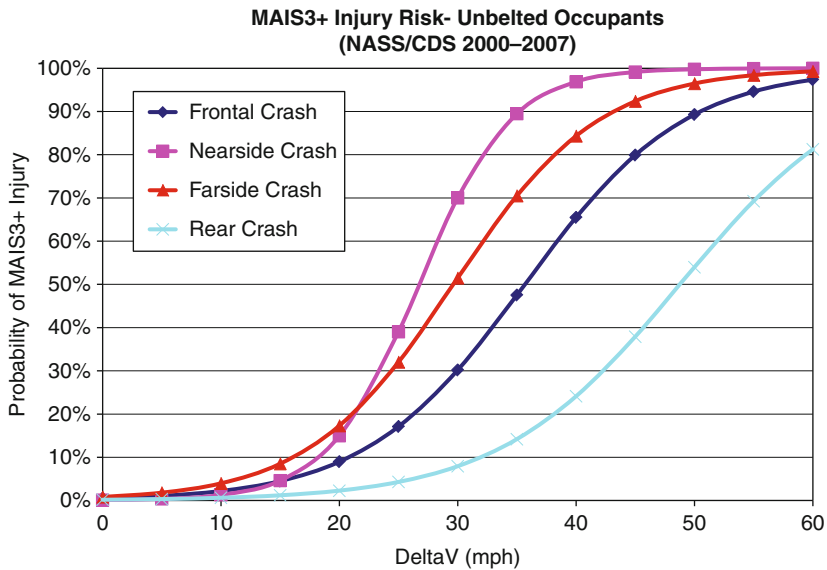
4 Injury Risk Threshold Selection

A principal aim in using eACN system data is to identify crash events where time critical injuries may have occurred so that rescue can be expedited. Once on the scene, rescue personnel can use additional criteria including physiological information to make decisions on whether or not an injured person needs to be transported to a Level 1 trauma center. As such, the eACN data will not form the sole basis for triage decision making.



■ Fig. 32.4

MAIS3+ injury risk by crash direction for belted occupants

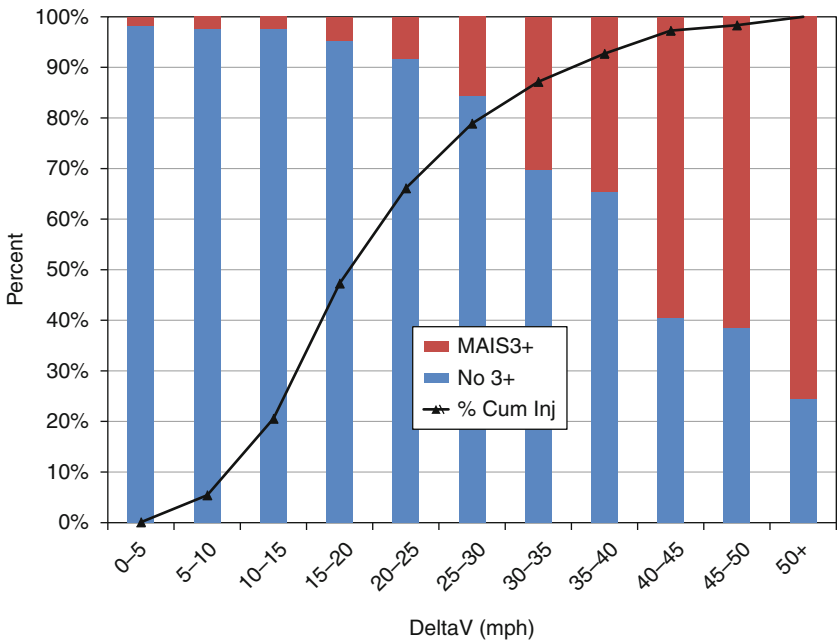


■ Fig. 32.5

MAIS3+ injury risk by crash direction for unbelted occupants

■ Table 32.3
URGENCY algorithm capture rate within the 2007 NASS CDS planar crash population

Crash mode	Sensitivity	Specificity
Frontal	71.2%	90.2%
Nearside	90.6%	85.7%
Farside	81.2%	88.6%
Rear	52.7%	98.2%
Overall	75.9%	90.8%



■ Fig. 32.6
Injured versus uninjured US Tow-away crash populations and cumulative percent by primary deltaV category

However, in the absence of additional data, these data should be considered reliable for making dispatch decisions.

Within the US crash population and around the world, the number of noninjured occupants far outweighs the number of seriously injured occupants at all crash severities up to 40 mph deltaV. While this trend highlights the excellent performance of vehicle structures, restraint systems, and other safety devices, it poses a difficult challenge for the identification of potentially injured occupants especially in the lower deltaV categories.

► Fig. 32.6 shows the percent of MAIS3+ injured occupants by deltaV category. As shown, the population of injured in the first four categories (up to a 20 MPH deltaV) all show that

less than 5% of occupants exposed to crashes within each group are severely injured. It should be noted that this figure combines all crash modes such that higher risk events like nearside crashes would show a different distribution when separated.

Within the first four deltaV categories, the risk of injury may in fact be low (less than 5% under 20 MPH), yet more than half of the population of serious injuries occurs at or below this deltaV. This reflects the high incidence of crashes at lower speeds. In estimating injury risk, other explanatory factors like crash direction, safety belt usage, multiple impact events, and rollover help to better identify injury risk yet some variability exists even within identical crash subgroups. In order to compensate for this natural variability, additional explanatory variables must be added to the injury model or some latitude must be given within the rescue system to allow for some over- and undertriage.

With this approach, a lower injury severity threshold is justified so that people with time critical injuries can receive rapid on-scene assistance. However, a large amount of overtriage due to false indication of injury would result in mistrust of the information by the rescue community. Overall, when a threshold of 10% is utilized (as described above), the rate of undertriage is minimized; however, this study indicates that the rate of overtriage is considerably large. The overtriage rate for a 10% risk exceeds the acceptable level for Trauma Center transport decision making but does not exceed current standards for EMS dispatch. When an alternative injury outcome is used ($ISS > 15$) and a 20% risk criteria is applied, a small percentage of the seriously injured population is correctly identified. Conversely, an MAIS3+, 10% risk criteria greatly improves the sensitivity of the model to detect the injured, yet this approach increases the rate of overtriage for the crash population evaluated.

Estimating the risk of ISS 15 or ISS 16 injury targets a very small percentage of the most severely injured population. Yet, injuries at ISS of 9 or even less often require medical care rapidly too (i.e., loss of consciousness, fractured bones with risk of serious internal bleeding – pelvic fractures, and thoracic injuries that compromise breathing and circulation). The ISS 9 and higher population is 5% of occupants traveling in vehicles 5 years old and newer involved in crashes severe enough to trigger the eACN system.

While one goal of eACN is to assist in subjective trauma triage decision making on scene, the limited nature of the vehicle sensor data may preclude trauma triage decisions remotely using sensor information alone. Rather, vehicle-based data should be relied upon to rapidly inform rescue of a potentially serious event where EMS can conduct on-scene evaluations to better establish the need for trauma center, Emergency Department, or other medical treatment when warranted. A lower injury risk threshold is more appropriate for rescue decisions than triage decisions.

A Center for Disease Control (CDC) Expert Panel recommended that triage to a trauma center should be required if the risk of ISS 15 or higher injury exceeds 20% based on crash factors including impact direction, crash deltaV, number of impact events, use or nonuse of safety belts, and the occurrence of rollover. A 10% risk threshold for MAIS3+ injuries is also used by some manufacturers to more conservatively identify those in need of medical attention at the scene of a crash but who may not necessarily require subsequent trauma center transport.

5 Future Opportunities to Improve Post-crash Care

Currently, a growing list of vehicle manufacturers offer ACN and eACN technologies in their vehicles in order to provide an enhanced level of safety and convenience in the event of a crash. To most effectively make use of this reliable data, a number of critical changes in dispatch, on-scene triage decision making, and in-hospital care are required.

Improved Methods to Exchange and Apply Crash Information During Dispatch: While TSPs are now capable of transmitting electronic data regarding a crash event directly to PSAPs (i.e., GPS coordinates), no information is supplied regarding the apparent severity of the crash or the likelihood of injury. In some cases, such data may be transferred verbally between the TSP and 911 operators. In order to promote widespread use of such data within PSAPs, dispatch protocols must be adapted to offer a standard treatment of telematics data including injury risk for any crash where it is known regardless of vehicle manufacturer. It is critical that the automotive industry reaches a consensus on how the data is treated and interpreted so that PSAPs receive an injury risk that is recognizable. A number of ongoing pilot studies funded by private industry and government organizations are underway to determine the most effective method for interpretation and transmission of such data and to understand the value of the data as perceived by 911 centers. Once complete, these studies will provide a foundation for more rapid adoption and use of the data by PSAPs.

Implementation of Enhanced Trauma Criteria: Although the US CDC has proposed the use of vehicle telemetry data as a criteria for a high-risk crash event, EMS staff and their trauma systems have not established best practices to treat or interpret vehicle data and they have not mandated its use during trauma triage decision making. Future policy changes are required so that the rescue community makes best use of this valuable information.

Improving Injury Risk Models: Currently, crash factors that are used to estimate injury severity include impact direction, crash deltaV per event, occupant seating position, use of safety belts, number of impacts, and the occurrence of rollover. Additional factors including occupant age, gender, and health status directly relate to injury tolerance as well. These occupant-specific factors may explain a portion of the variability in injury outcome, yet such information is difficult to detect using in-vehicle technologies alone. In the future, methods to identify occupants verbally or using in-vehicle technologies will greatly improve the accuracy of injury models through the inclusion of occupant-specific factors. Such factors may be stored remotely for occupants who choose to make such data available in the event of an emergency.

Use of Enhanced Data Within Hospitals and Trauma Centers: In order to improve patient care, clinicians require contextual awareness of the events leading to a patient's arrival in an emergency department or trauma center. By linking crash records collected by onboard vehicle systems and making this information available at the point of care, physicians may target evaluations once they have established a clear picture of a potential mechanism of injury. Further, improved patient history including preexisting conditions, medications taken, or other pertinent health data will allow for optimal care. Making such information available at the point of care may, in the future, become a reality as more rapid patient identification is made.

6 Conclusion

Advanced telematics data has the potential to substantially impact morbidity and mortality if appropriately applied during the post-crash phase. Similarly, the opportunity to remotely and instantaneously assess the overall severity of a crash event could also positively impact the appropriate allocation of rescue services. At the present time, vehicle manufacturers have developed and implemented sensor technology and communication systems to collect and transmit accurate crash metrics to describe the nature of a crash. To date, however, 911 and EMS services have not implemented protocols to capture, interpret, and act upon such information received electronically. This final step will require active participation of government and regulatory agencies, industry stakeholders, and the medical community to evaluate and accept such systems to improve outcome for crash victims.

References

- ACS (2006) Resources for optimal care of the injured patient 2006. R Coscia, J Meredith. American College of Surgeons Committee on Trauma, Chicago
- Bedard M, Guyatt G et al (2002) The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accid Anal Prev* 34(6):717–727
- Clark D, Cushing B (2002) Predicted effect of automatic crash notification on traffic mortality. *Accid Anal Prev* 34(4):507–513
- Cudnik M, Newgard C et al (2008) Distance impacts mortality in trauma patients with an intubation attempt. *Prehosp Emerg Care* 12(4):459–466
- Cummings P, Rivara F (2004) Car occupant death according to the restraint use of other occupants: a matched cohort study. *J Am Med Assoc* 291(3):343–349
- Cunningham P (1995) North Carolina preventable mortality study with inter-rater reliability modifications. NHTSA report DOT HS 808 345
- Esposito T, Illness C et al (1992) Rural preventable mortality study. National Highway Traffic Safety Administration; NTIS [distributor], Washington, DC
- Farmer C, Braver E et al (1997) Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity. *Accid Anal Prev* 29(3):399–406
- Jones I, Champion H (1989) Trauma triage: vehicle damage as an estimate of injury severity. *J Trauma* 29(5):646–653
- Kohn M, Hammel J et al (2004) Trauma team activation criteria as predictors of patient disposition from the emergency department. *Acad Emerg Med* 11(1):1–9
- MacKenzie E, Rivara F et al (2006) A national evaluation of the effect of trauma-center care on mortality. *N Engl J Med* 354(4):366–378
- Maio R, Burney R et al (1996) A study of preventable trauma mortality in rural Michigan. *J Trauma* 41(1):83–90
- NHTSA (2010) Traffic safety facts: highlights of 2009 motor vehicle crashes. National Center for Statistics and Analysis, NHTSA, Washington, DC
- Norwood S, McAuley C et al (2002) A prehospital Glasgow coma scale score ≤ 14 accurately predicts the need for full trauma team activation and patient hospitalization after motor vehicle collisions. *J Trauma* 53(3):503–507

33 Map Data for ADAS

John Craig
JCC, Munich, Germany

1	<i>Introduction: The Use of Maps Today Within the Automotive Industry</i>	882
2	<i>Energy-Management Applications Using Maps</i>	883
2.1	Fuel-Efficient Routing (Also Referred to as “Eco-routing”)	883
2.2	Electric Vehicle Range Prediction	884
2.3	Eco-driving/Predictive Gear-Shifting/Predictive Cruise Control	885
3	<i>Safety Applications Using Maps</i>	886
3.1	Predictive Front Lighting	886
3.2	Overtaking Assistant	887
3.3	Curve Speed Warning	888
4	<i>Requirements on Maps from These New Applications</i>	889
5	<i>Future Directions for Maps in Vehicles</i>	890
6	<i>Conclusions</i>	892

Abstract: From their beginnings in the car as a tool for A-to-B navigation, digital maps are experiencing an evolution process that will see them at the forefront of new applications designed to improve active safety and manage fuel consumption. These maps will, in effect, become a new vehicle sensor, with a range exceeding that of camera and radar systems, and an ability to work in all weathers and at night. These new maps will need to be more accurate than those used for navigation, and be fused with a minimized set of map attributes to create new vehicle-interpreted precision maps. This chapter will look at the applications that would benefit from these new maps, which in terms of both safety and energy management applications, provide precise knowledge of the road ahead. This allows the vehicles and drivers to be informed of potentially dangerous situations, and take actions based on exact knowledge of future slopes and curves in the road. In energy management terms, the knowledge of road slope will allow the most fuel-efficient routes to be chosen, and can be used to determine the range of Electric and Hybrid Electric Vehicles (EV/HEV), as well as optimizing engine and transmission for fuel efficiency. We will consider how such maps can be created using a number of different technologies, and how this collection methodology impacts their characteristics. As maps evolve and become more “connected,” the possibilities to update them, and access further geographic data and services, will further increase their usefulness.

1 Introduction: The Use of Maps Today Within the Automotive Industry

Digital maps have been around for some time, and their use for vehicle navigation is something that was pioneered by the automotive industry in the 1990s. Today their prime use is still in vehicle navigation systems, as a base onto which your current position can be projected, a destination selected, and an optimized route given for the driver to follow both visually and audibly. Later in this section, we will discuss uses of maps other than simply guiding us from A to B, but for now navigation is the overwhelmingly dominant use.

Having a map in a car gives that vehicle knowledge of its surroundings. It can be thought of as an additional sensor, alongside others such as camera and radar systems. Compared to them it usually lacks a real-time element and does not provide as rich information in the vicinity of the vehicle. On the other hand it effectively has unlimited range, and is not susceptible to poor weather, poor lighting, or road conditions.

What can this map tell us, as a driver? Generally, a basic navigation map gives the approximate shapes of the roads, the size of roads, the use of land around these, and further topological information. More advanced maps can indicate information such as current speed limits, nearest Points-Of-Interest (POI – examples include hotels and restaurants), and more. With real-time connectivity, the same basic map can support information such as traffic, weather, and even concepts extending into where all your friends are located. Going further, we can see a use case whereby there is no selected route, but the driver has a map view of all static and dynamic data around them. This allows them to check quickly the traffic situation, nearest burger bar, or whatever, depending on their preferences.

Therefore it is apparent that maps can support both A to B navigation, and a more free-form helicopter view of items the user has decided are of relevance around them. If we extend the functionality of maps even further, imagine if a car had detailed knowledge of the road ahead, before driving on that road. It already knew every curve, incline, dip and crest, of every road in the country, to a high accuracy, and in three dimensions. With this information and knowledge, what could the car then do?

In the next section we will explore these possibilities, through which the map acts as a vehicle sensor in support of new applications.

2 Energy-Management Applications Using Maps

How can maps make vehicles more fuel-efficient? Here we describe some application examples where the use of map data can directly lead to reduced fuel consumption and less harmful emissions (Li and Tennant 2009).

2.1 Fuel-Efficient Routing (Also Referred to as “Eco-routing”)

With an increasing number of vehicle kilometers driven per year there is a corresponding increase in fuel usage, taking up a large percentage of energy use in a country and causing harmful emissions. More fuel-efficient transport systems can reduce fuel costs and dependencies, and limit environmental damage. One solution in this direction is for drivers to select routes which are the most fuel-efficient, taking all relevant parameters into account.

Navigation systems already balance distance and time, and route via a number of different methods. The shortest route will often take smaller country roads and neighborhood streets. These are generally slow routes, with many starts and stops, and result in high fuel consumption. Using the fastest route normally means selecting highways and ring roads, which means fewer intersections and less start and stops, but longer distances and thus also implying high fuel consumption. A third choice is the most fuel-efficient route (and lowest harmful emissions) which requires that fuel consumption costs per road segment or link, are known and used by the navigation system routing algorithms to select the best route.

The most fuel-efficient route has to balance a number of parameters which impact the fuel economy, some of which are route-dependent, and some are not, as show in the diagram below (► Fig. 33.1).

The obvious parameters which are route dependent will include:

- Road slope: Going uphill consumes much more fuel, and generally cannot be offset by downslopes. Thus, the amount and severity of upslopes in a route is a key consideration.
- Road junctions, pedestrian crossings: Stop and go driving consumes more fuel.
- Speed limits: There are optimum speeds for vehicle fuel economy.

Mainly route-dependent	Mainly route - independent
Static Parameters	Vehicle Parameters
Road slope	Engine type and efficiency
Road junctions	Transmission type
Speed limits	Size and Weight
Pedestrian crossings	Drag coefficient
Speed bumps	Other electronic systems (on/off)
Historical Traffic Data	Regenerative breaking
Dynamic Parameters	Driver Parameters
Real-time Traffic Data	Driving style
Weather	
Road rolling resistance	

■ Fig. 33.1
Parameters relevant for fuel-efficient routing (Source: Intermap)

- Speed bumps: Slowing down and accelerating consumes more fuel.
- Historical Traffic Data can predict the speed and degree of stop–start driving.
- Road rolling resistance determines inefficiencies connected to the road surface.
- Real-time traffic indicates current slow or stopped areas on any route.
- Weather, including wind direction and strength, can have an impact.

A typical eco-routing algorithm will take these factors into account in calculating the most fuel-efficient route. Map data will form the basis and source for these parameters. To the driver it will look like any other route, but will mean that they use less fuel than any alternative routings. Studies have shown (Boriboonsomsin and Barth 2009) that a hilly route compared to a flat route can consume 15–20% more fuel. Thus information on road slope is required, and can be preassigned to road segments such that routing algorithms can process this data and assess routes according to slope severity and frequency. Slope and other data allow the route which uses the least fuel to be selected by the navigation system.

2.2 Electric Vehicle Range Prediction

This is a derivative of the previous Eco-Routing example, with some modifications. Users of Electric Vehicles (EV) can suffer from “range anxiety” and not knowing precisely when they will run out of battery power and left stranded by the roadside. Driving an Electric Vehicle is sometimes comparable to driving a combustion-engined vehicle that has the low fuel indicator perpetually lit. The range of an EV is currently limited by battery technology, and they can be more sensitive to parameters that impact energy consumption than similar internal combustion-engined cars. It is therefore important that the range of EVs can be accurately and reliably predicted, taking all relevant factors into consideration.

This usually manifests itself in a map-based range diagram, showing the furthest point a vehicle can reach on roads in all directions, and also the point of no return along such roads. In order to judge how far along a particular road the EV can travel, knowledge of the road slope and other features is required.

Through having confidence in the range prediction, drivers will become trained to adopt more efficient charging patterns, prolong battery life, and perhaps even facilitate smaller batteries. In effect, the detailed slope knowledge and other data are used to more actively management the vehicle state of charge.

Due to their limited range and charging options, owners of EVs need map-based driver assistance systems for information such as:

- Charging Station locations
- Nearest Charging Station occupancy status
- Range left on battery in all directions
- Point of no return calculations in all directions

These requirements may well imply that a form of navigation and telematics system may be mandatory for such vehicles, for them to function properly.

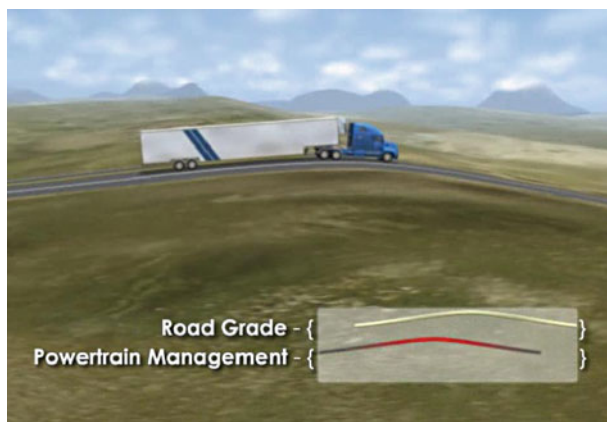
An application related to this, is the management of Hybrid Electric Vehicle charge and discharge cycles (Zhang et al. 2009a, b). For example, knowledge of an upcoming steep hill through GPS and 3D road maps will be a hint to the power management scheme to charge the batteries up in anticipation of larger power demand during the uphill ascent. Similarly, the battery can be discharged before a downhill descent in anticipation of excess power becoming available when going downhill.

2.3 Eco-driving/Predictive Gear-Shifting/Predictive Cruise Control

Although sounding similar, this is a somewhat different Application to those previously discussed, and is today mainly applicable to the Commercial Vehicle segment. Eco-Driving is a function whereby the vehicle uses map-derived knowledge of the road slope ahead, to optimize the gear and engine/cruise control systems for lowest fuel consumption. In advanced systems, the transmission or cruise control systems adapt automatically to upcoming slopes (► Fig. 33.2).

The fuel savings are achieved through selecting the optimal gearing or acceleration in anticipation of road slope changes indicated by this new generation of highly accurate road geometry maps. When the vehicle knows that a downhill slope is ahead, it can avoid unnecessary acceleration. With an uphill slope ahead it can build momentum. Knowing that the road ahead is relatively flat can also permit a degree of coasting by the vehicle, using its momentum to propel the vehicle while in neutral. The key is ensuring that inflection points where the road slope changes are accurately mapped. Such Eco-Driving systems will save the truck owner some percentage of fuel costs per year, which can accumulate into large amounts across truck fleets and can justify such an investment.

Other Eco-Driving systems are less integrated and can be brought into the vehicle with knowledge of a particular vehicles engine and transmission characteristics, and just offer the driver advice on how to achieve the best fuel efficiency through changing gear,



■ Fig. 33.2

Upcoming road slope used by powertrain management (Source: Intermap)

braking, driving at optimal speeds, and more. The driver can subsequently get feedback on how efficiently they have driven.

3 Safety Applications Using Maps

The next generation of accurate road maps can also make vehicles safer and help drivers avoid accidents, through either advance warning off difficult situations, or by bringing a more predictive element to existing systems. Below are a few examples where map data is used to help the driver avoid dangerous situations.

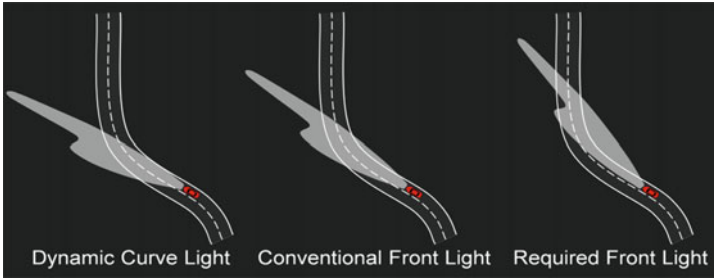
3.1 Predictive Front Lighting

There are front lighting systems today that follow the motion of the steering wheel and turn into corners. While generally an improvement there are situations, such as S-bends, whereby such dynamic curve lighting can be even worse than conventional lighting (🔗 Fig. 33.3).

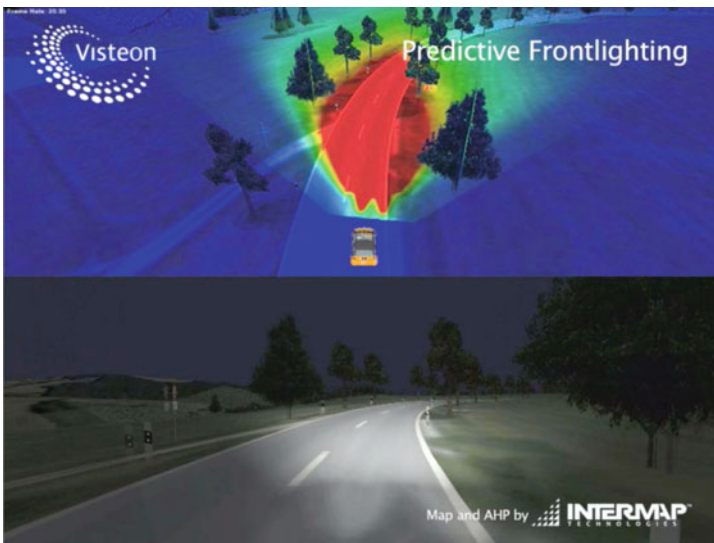
There are also systems that know when they are on major highways, or in villages or urban areas, and modify their headlight beam accordingly. This is the beginning, but such systems can be developed further, for very good reasons.

According to an American study, the risk of accidents with pedestrians and game animals rises by the factor 4 at night. In Germany, 42% of all fatalities happen at night. Driving at night on roads with a high proportion of curves, dips, slopes, and other features, presents the highest potential for danger.

Headlights that can anticipate such situations have value and have been built and researched by many companies. Using accurate road geometries the vehicle headlights



■ Fig. 33.3
Lighting systems in an S-bend (Source: Intermap)



■ Fig. 33.4
Using maps to direct headlights through upcoming bends (Source: Intermap)

possess knowledge of the road ahead, which is not impacted by weather, lighting, or road surface considerations. Naturally the map would not be used as the only sensor, but would be combined with for example, a steering wheel sensor system, or camera system. Results have shown that map-based systems can offer optimal illumination on the road ahead, and are significantly better than non-predictive systems, or those using only steering wheel sensors (► Fig. 33.4).

3.2 Overtaking Assistant

This can be an extremely useful assistant to the driver, but also needs to be carefully designed in terms of the advice it gives. Overtaking is a significant source of accidents. As with predictive front lighting, the most challenging situations occur on smaller roads,

in more rural areas, with roads featuring significant curves, dips, and slopes, with added dangers at night. Technologies and data are available today, which could provide driver information, which could reduce overtaking accidents and be built into an “Overtaking Assistant.” The variables of relevance in this case include:

- Line-Of-Sight (LOS), a computation based on
 - Road curvature and knowledge of upcoming bends
 - Road slope and knowledge of dips and crests obscuring oncoming vehicles
 - Surrounding Terrain knowledge in three dimensions
- Distance to Junctions and other road features requiring care to be taken
- Current speed and current speed limit
- Historical traffic data (probability of encountering another car)
- No overtaking zones
- Accelerative capability of vehicle (though this cannot be assumed to be fully used)
- Driver height above road
- Radar information
- Camera information
- Visual information and, in particular, is there an oncoming vehicle now

Of particular interest from a mapping perspective is the availability of data allowing an LOS attribute to be computed. This would indicate a distance or time where the driver can clearly see the road ahead, with the greater the time/distance implying the safer it would be to overtake.

Without going deeply into legal aspects, it is simpler to offer advice on when not to overtake, due to any of the above parameters indicating a danger. Advising a driver at any point that it is OK to overtake carries certain liabilities. A useful application would be when a driver is stuck behind a slow-moving vehicle, which can cause frustration as the driver is continually looking for overtaking opportunities. If there was a means to inform the driver that they should wait for a certain time or distance until a much better opportunity arises, then they could relax until this appeared. Technically this is feasible; the challenge is how to inform the driver and accommodate legal aspects.

3.3 Curve Speed Warning

Curve speed warning (CSW) helps drivers identify potentially dangerous situations if a bend in the road is taken too fast, and warns the driver in advance allowing him time to react properly. The information about such bends is drawn from digital maps of the road and analysis of the geometric characteristics of the bend. By optionally combining this information with external factors such as weather conditions, the maximum recommended speed for the bend is estimated. If the vehicle is approaching at a speed higher than the recommended value, the system can warn the driver of the potential hazard, prepare the safety systems in the vehicle, or actively inhibit further acceleration of

the vehicle. Traditionally implemented only in the horizontal plane (XY), there is added value in including the vertical (Z) plane, given that it is easier to reduce speed on an uphill curve and downhill curves are inherently more dangerous.

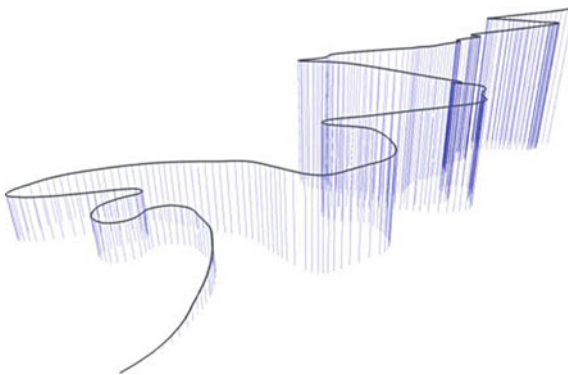
Curve speed warning is also useful for Commercial Vehicles, where accidents still occur due to vehicles rolling or tipping over. Precise curvature and slope information can be fused with information from, for example, tire pressure sensors, load weight measurements, road surface information, and more, to alert drivers to such dangers and cause a reduction of such fatalities.

4 Requirements on Maps from These New Applications

Indicating to a navigation system user that you now intend to use these maps for safety systems, may not always solicit a positive reaction. Maps in navigation systems tend to be fit-for-purpose (A to B routing) and users will overlook aspects like missing streets, wrongly placed streets, routing errors, and other anomalies – because the system usually gets it right. However, when moving to energy management and safety applications, such maps have limited value and there is a need for this new generation of more accurate maps (Dobson 2009).

First and foremost is the need for a new way of defining a road, more accurately and in three dimensions. These road models or maps, sometimes known as ADAS road geometries, are essential to support the applications described and are available from several suppliers.

The model comprises the standard horizontal (XY) road shape, with the addition of height information along the road (● Fig. 33.5). These topological road models are described as a series of connected nodes and segments or vectors. One way of representing the road is to define points along the centerline at variable intervals proportionate to the spacing required to accurately define road shape. That is, tighter curves may need points close together while longer straights can use much less points and still capture the road shape. More useable



■ Fig. 33.5

Three-dimensional road model (Source: Intermap)

formats can be created, such as road slope expressed as a percentage incline over a fixed segment length, curve radii, or mathematical formulas such as splines and clothoids which approximate the smooth road shape. Specific attributes can also be associated with this road model, such as the road function class, and the existence of bridges and tunnels.

A key distinction from the roads in navigation systems, is that these more accurate ADAS geometries are usually machine-interpreted and without a human-machine interface. They can be handled directly by the applications described above, and in some cases are processed via an Electronic Horizon (eHorizon) system. Such systems use the vehicle position to provide the upcoming road geometries to a vehicle application.

These accurate road geometries can be created by a variety of different means, with the three most prevalent being to either:

1. Drive the roads with a fleet of suitably equipped *mapping vans*. These vans are equipped with an array of technologies including sophisticated cameras, GPS, inertial measurement units, lasers, lidar and other technologies. The vans are driven along the roads and capture all relevant data as they go.
2. Fly over the roads with a radar-equipped plane, known as *remote sensing*. This can involve dual radar installations on small jets which go up and countries in strips to ensure full coverage. Such radar-based systems can operate at night and in bad weather, to ensure more rapid collection.
3. Process large numbers of reports from GPS devices (“probes”) that have driven the roads and transmitted their (XYZ) coordinates back to a server. The data is generated and transmitted automatically in huge quantities by suitably equipped PNDs and Phones, and processed using complex algorithms to derive road geometry.

Each method has generic advantages and disadvantages (● [Fig. 33.6](#)).

The most suitable method for any application is dependent on the requirements on cost, coverage, and accuracy.

5 Future Directions for Maps in Vehicles

We have described the current use of maps in vehicles for navigation, and some new applications that are currently being worked upon, and are either in Research and Development, Pre-Development, or Series production in some cases. If we look either further ahead, or take a broader perspective on such technology, there are even more interesting applications on their way.

With accurate road geometries it may be possible to have a further independent positioning method, whereby the camera view of the road geometry is matched to the map, and a position generated. Such methods could be used to secure autonomous driving, or support augmented reality navigation where the camera view has navigation information superimposed on it. GPS positioning can effectively be confirmed or refined using such methodology.

	Road data collected from Remote Sensing	Road data collected from Driving	Road data collected from Probes
Speed of Collection	Fast	Slow	Medium
Accuracy	To be determined. Little independent public information is available. In terms of what is required, different applications and car manufacturers will demand different accuracies		
Cost	Lower	Reference	Lowest
Coverage	100% of roads per country covered from day 1	Built up gradually by driving major roads first	Built up gradually by roads with most traffic = probe reports
Other Issues	Can be difficult to sense roads in areas of obstruction & high slope. Auxiliary data required.	To collect all roads takes a long time and has high costs	Needs to gain industry acceptance. Not enough reports to create accurate geometries for quieter roads

■ Fig. 33.6

Comparison of methods to derive 3D road geometries (Source: Intermap)

One key area of development is sensor fusion, whereby existing physical sensors built into vehicles are fused together with map information to improve the overall function or feature. Often such sensors serve a single purpose and the challenge will be to use them for more generic purposes, together with maps. Combining sensors in this manner can make the overall vehicle more robust and reliable.

As fast as cars evolve, consumer devices evolve faster. Thus the devices brought into cars, such as Phones and PNDs, will have a tendency to be significantly more advanced than those already embedded there. The question is whether these nomadic devices can be leveraged for some of the energy management and safety applications discussed previously. Could accurate road geometries be brought into the vehicle and function just as well as if they were integrated during production? ([Craig](#)).

A further trend will be the proliferation of reliable vehicle connectivity whereby, for example, applications and data can be updated or even executed on a server, and communicated to the vehicle. This can allow for handling of greater amounts of information, and shorter development cycles without the need to have everything embedded.

The use of new technology in vehicles continues at a high pace, and the use of maps will be important to some of the key automotive developments in the near future.

6 Conclusions

We have briefly described some uses of accurate 3D road geometries, and others will undoubtedly emerge over the coming years. It is clear that better quality map data will mean that it can be used beyond the traditional realms of navigation or simple browsing. These new maps will not be designed for human interpretation, and therefore much of the heavy visual aspects can be removed from the data and associated software. Instead, the geometries and some limited attributes will form new and simpler “maps for cars,” interpreted by the vehicle to enhance occupants safety and reduce environmental damage through improved fuel management.

References

- Boriboonsomsin K, Barth M (2009) Impacts of road grade on fuel consumption and carbon dioxide emissions evidenced by use of advanced navigation systems. *Transport Res Rec* 2139:21–30
- Craig J There will be maps for cars, and maps for people. *GPS Business News*. http://www.gpsbusinessnews.com/There-Will-Be-Maps-for-People-and-Maps-for-Cars_a2011.html
- Dobson MW (2009) Creating robust functionalities, ADAS and 3D-road map databases. *GeoInformatics* 28–33
- Huang W, Bevly DM, Li X, Schnick S (2007) 3D road geometry based optimal truck fuel economy. In: *Proceedings of ASME international mechanical engineering congress and exposition*, Seattle, Washington, 11–15 Nov 2007
- Li X, Tennant K (2009) Vehicle energy management optimization using look-ahead three-dimensional digital road geometry. In: *ITC World Congress*, Stockholm, The Netherlands, 21–25 Sept 2009
- Zhang C, Vahidi A, Li x, Essenmacher d (2009a) Role of trip information preview in fuel economy of plug-in hybrid vehicles. In: *ASME 2009 dynamic systems and control conference (DSCC2009)*, Hollywood, California, 12–14 Oct 2009
- Zhang C, Vahidi A, Pisu P, Li X, Tennant K (2009b) Utilizing road grade preview for increasing fuel economy of hybrid vehicles. In: *Proceedings of the 12th IFAC symposium on control in transportation systems*, Redondo Beach, February 2009

Drowsy and Fatigued Driver Detection, Monitoring, Warning

Azim Eskandarian

34 Advances in Drowsy Driver Assistance Systems Through Data Fusion

Darrell S. Bowman · William A. Schaudt · Richard J. Hanowski
Center for Truck and Bus Safety, Virginia Tech Transportation
Institute, Transportation Research Plaza, Blacksburg, VA, USA

1	<i>Introduction</i>	896
2	<i>Defining the Problem: Fatigue Versus Drowsiness</i>	897
3	<i>Salient Measures of Driver Drowsiness</i>	897
3.1	Driver-Based Approaches	898
3.1.1	Electroencephalography (EEG) Measure	898
3.1.2	Ocular Measures	898
3.2	Vehicle-Based Approaches	900
3.2.1	Lane Position/Line Crossing	900
3.2.2	Steering Wheel Inputs	900
3.3	Measures of Driver Drowsiness Summary	901
4	<i>Development of a Robust Drowsy Driver Assistance System</i>	901
4.1	Case Study of Prototype DDAS Utilizing a Data Fusion Approach	904
4.1.1	Drowsiness Indicator Selection	905
4.1.2	Fusion of Data from Two Drowsiness Indicators	906
4.1.3	Testing and Evaluation of Prototype DDAS	908
5	<i>Conclusion</i>	909

Abstract: Every year, thousands of vehicles are involved in crashes which are attributed to the onset of driver drowsiness. As a result, there are numerous drowsy driver assistance systems (DDAS) available on the market; however, many of these technologies rely on a single predictor of driver drowsiness (e.g., eye closures, lane position, steering). Relying on only one predictor of drowsiness makes the system susceptible to periodic intervals in which data is unavailable due to failure of the single sensor, usage outside of the sensor's envelope of operation, or driver's individual differences. Driver drowsiness measures can be classified as either driver-based (those measures derived from the human) or vehicle-based (those measures derived from the vehicle). For driver-based measures, PERCLOS (a measure of slow eye closure) is considered to be a robust measure of driver drowsiness. Machine-vision (MV) slow eye-closure sensors have been developed to estimate the percent of eye closure and calculate the PERCLOS value. However, these MV slow eye-closure sensors' ability to detect the eye closures is challenged by eyewear, ambient illumination, and head movement. For vehicle-based drowsiness metrics, lane position appears to be a key indicator of driver drowsiness. Lane position is typically estimated through MV technology detecting lane edge markings on the forward roadway scene. The absence of lane edge markings on roadways or instances of low contrast between lane markings and the surrounding scene make it difficult for this MV lane position sensing technology to accurately measure the vehicle's position within the lane. Typical causes of low contrast lane markings include poor lane marking quality, artificial overhead lighting, or headlight "blooming." Therefore, a multi-measure approach, that uses multiple distinct sensors, can offer not only sensor redundancy but also provide a data fusion approach whereby both measures provide more robust drowsiness detection than either measure could alone. This chapter describes the salient measures of driver drowsiness, the concept of data fusion in DDASs, and provides a case study of a prototype DDAS that integrates two drowsiness metrics (i.e., PERCLOS and Lane Position) to form an enhanced drowsiness estimate that may prove to be a more robust measure in a real-world, field application as compared to a single metric system.

1 Introduction

Driver impairment due to drowsiness is known to be a major contributing factor in many crashes. For example, several sources indicate driver fatigue is the probable cause of approximately 30% of crashes (National Transportation Safety Board 1990; Kecklund and Akerstedt 1993; Horne and Reyner 1995; Folkard 1997; Lenne et al. 1997; Lyznicki et al. 1998). Given the impact of driver drowsiness on driving safety, there has been a keen interest in developing safety systems that can monitor and quantify driver drowsiness and provide a real-time warning to the driver and/or a control output to the vehicle or other systems as warranted. There are numerous drowsy driver monitoring systems available on the market; however, nearly all of these technologies rely on a single predictor of driver drowsiness (e.g., eye closures, lane position, steering). One downside of using only one

predictor of drowsiness is that they are susceptible to periodic intervals in which data is unavailable due to failures of the single sensor, operation outside of the sensors envelope of operation, or driver's individual differences (e.g., eyewear or skin tone). When the single predictor does not work optimally, the system is less effective than it could be. Therefore, a multi-measure approach, that uses multiple distinct sensors, can provide not only a backup system with sensor redundancy but also provide a data fusion approach whereby both measures provide more robust drowsiness detection than either measure could alone.

2 Defining the Problem: Fatigue Versus Drowsiness

It is important to note that the terms *fatigue* and *drowsiness* are often used interchangeably in the literature. This chapter will use the terms interchangeably in the manner in which they are presented in the original literature sources. This is to ensure that the authors' operational definitions and intentions remain unchanged. However, a distinction between the terms is made at times, and this distinction is evident by comparing the definitions below.

Fatigue is defined as "a state of reduced physical or mental alertness which impairs performance" (Williamson et al. 1996, p. 709). Another definition provided by Dinges (1995) is "a neurobiological process directly related to the circadian pacemaker in the brain and to the biological sleep need of the individual." Dinges further states that fatigue is something all humans experience, noting it cannot be prevented by any "known characteristics of personality, intelligence, education, training, skill, compensation, motivation, physical size, strength, attractiveness, or professionalism" (1995, p. 42).

Drowsiness is defined as the "inclination to sleep" (Stutts et al. 1999) and is also commonly referred to as "sleepiness." As noted above, fatigue is a reduced state of mental or physical alertness that impairs performance. Fatigue can occur without actually being drowsy; therefore, *fatigue* and *drowsiness* are not exactly synonymous. Where fatigue is the result of physical or mental exertion, drowsiness may result from boredom, lack of sleep, hunger, or other factors.

3 Salient Measures of Driver Drowsiness

Quantifying driver drowsiness is an inexact, lagging measurement process. Unlike other driver-impaired states, such as alcohol or drug intoxication, drowsiness cannot be immediately measured with a breath or urine sample. Also, the onset of fatigue or drowsiness is not an instantaneous occurrence, but a cumulative process. Drowsiness takes considerable time, often an hour or more, before it manifests with noticeable differences in a driver's physical characteristics, performance, or mannerisms (Knipling and Wierwille 1994).

Recent advances in drowsy monitoring technology have demonstrated the ability to detect potentially dangerous levels of drowsiness through a range of different approaches. These approaches include (1) driver-based measures of drivers' current state of well-being ascertained by the physiological changes in EEG, pupil occlusion, ocular movements, etc., (2) vehicle-based measures of driving performance using variables such as speed, lateral lane position, time-to-line crossing, lane drift, steering movements, etc., and (3) a combination of person- and vehicle-based measures. Continued research is attempting to determine which of these salient indicators are most important to accurately and reliably measure drowsiness.

3.1 Driver-Based Approaches

3.1.1 Electroencephalography (EEG) Measure

While there are numerous physiological indicators to measure the level of alertness, the EEG signal is considered a more reliable and predictive means for measuring alertness levels (Erwin 1976; Volow and Erwin 1973; Artaud et al. 1994). Owing to this finding, the EEG has been regarded as the standard for measuring alertness/drowsiness both in the laboratory and some limited use in field studies (Brookhuis et al. 1986; Torsvall and Åkerstedt 1987; Brookhuis 1995; Wylie et al. 1996).

There are two important limitations of the measure to be considered. The first is data artifacts such as coughing, sneezing, vibration, and large body movements. In a simulator study using an EEG-based algorithm to detect driver fatigue, Lal et al. (2003) stated that additional research is needed to produce a real-time, robust, and reliable fatigue-detecting/alerting system that minimizes these data artifacts. The second limitation is this measure's intrusiveness to the driver. Today's EEG technology requires the driver to wear a skullcap, which is connected to a computer by wires and may be constraining and uncomfortable. While wireless EEG systems are in development, they still require the individual to wear head-mounted electrodes.

3.1.2 Ocular Measures

Ocular measures are commonly in the form of eye closures, eye blinks, pupil responses, or other eye movements. Bhuiyan (2009, p. 418) noted four basic steps employed by real-time, in-vehicle video-based systems:

1. Localization of the eyes (in the first frame)
2. Tracking the eyes in the subsequent frame
3. Detection of success or failure in tracking
4. Detection of possible drowsiness based on pre-defined algorithm logic

There are two primary methods of measuring ocular movements; namely, slow eye closures and eye blinks. While an eye blink is typically a very quick closure and re-opening of the eyes, “slow eye closures” are relatively gradual eye movements where the eyelids droop and close slowly.

In the mid-1970s, Erwin et al. (1973) and Erwin (1976) found that slow eye movements provided a valid means to measuring drowsiness levels in both narcoleptic and normal subjects. This work was continued in the 1980s, when Skipper and Wierwille (1985, 1986) defined and tested a variety of measures of slow eye closures (i.e., PERCLOS – the proportion of time interval that the eyes are 80–100% closed; AVECLOS – the sample mean of eye closures; and EYEMEAN – the sample mean of the square of eye closures) in a motion-based driving simulator. Their findings included moderate correlations between the various slow eye-closure measures and lane-related measures (e.g., range of $r = .50$ to $.60$ for lane deviation). Of the numerous slow eye-closure measures, estimated PERCLOS has been recognized as the most effective ocular measure for detecting the onset of driver drowsiness (Skipper and Wierwille 1986; Knipling 1998; Bhuiyan 2009). In fact, estimated PERCLOS is the only ocular measure to be validated by the National Highway Traffic Safety Administration (NTSB 1990; Dinges and Grace 1998).

The spontaneous eye blink has also been examined as an ocular indicator of fatigue (Stern et al. 1994; Caffier et al. 2003; Åkerstedt et al. 2005). Typically, researchers use metrics such as fixation duration, saccadic velocity, blink frequency (or rate), blink amplitude, blink duration/interval, and lid-closure velocity to describe eye blinking behaviors. For example, Picot et al. (2010) explored the algorithm for drowsiness detection based on visual signs that were extracted from high-frame-rate video. They proposed and tested an algorithm that merged the most relevant blinking features (duration, PERCLOS, frequency of the blinks and amplitude-velocity ratio) using fuzzy logic. They tested the algorithm on a data set representing 60 h of driving from 20 different drivers and found 80% good detections of drowsy states (Picot et al. 2010).

A primary confound with estimating drowsiness with eye blinks are the distinct differences among individuals’ eye blink behavior as it relates to drowsiness. Ingre et al. (2006) noted that individual differences are complex and the underlying principles behind them are not yet understood. It is clear that more research is needed to address these individual differences before an effective Drowsy Driver Assistance Systems (DDAS) can fully be developed based upon these aforementioned specifications. At this point in time, it would seem PERCLOS would be a more suitable indicator of fatigue than eye-blink behaviors.

Both of these ocular-based measures of drowsiness (i.e., slow eye movements and eye blinks) can be measured in a noninvasive manner by directing infrared (IR) light toward the eye which is then reflected back from the eye and recorded using a video camera. This technique may work effectively in ideal low-light conditions but may fail during daylight or well-lit road conditions (Wierwille et al. 2003; Bhuiyan 2009); as well as with the presence of eyewear such as prescription glasses or sunglasses and rapid changes in head position.

3.2 Vehicle-Based Approaches

3.2.1 Lane Position/Line Crossing

Measures of lane position have been found to be a capable indicator of driver drowsiness (Skipper and Wierwille 1986; O’Hanlon and Kelly 1974; Chatterjee et al. 1994; Wierwille 1994; Tijerina et al. 1999). Lane position/line crossing research appears to provide strong evidence that lane keeping degrades as a function of drowsiness impairment. The following lane metrics have been found to be measures of degraded driving performance because of fatigue:

- Lane-keeping/lane-tracking capability,
- Lane-drift frequency (standard deviation of lane position),
- Line crossing, and
- Time-to-lane crossing (TLC).

It is unclear which metric of lane tracking (e.g., lane position, mean square of lane position, variance of lateral position, standard deviation of lateral position relative to lane, proportion of time that any part of the vehicle exceeds the lane boundary, mean square of the difference between the outside edge of the vehicle and the lane edge when the vehicle exceeds the lane, variance of the time derivative of lane position, and standard deviation of the time derivative of lane position) is the most suitable metric. Nonetheless, lane-related behaviors demonstrate promise as a salient drowsy driver indicator.

As with slow eye closures, Lane position/line crossing is not a completely robust drowsiness measure. For example, lane position data loss can occur when the system fails to read the lane’s edge markings due to (1) weather (e.g., rain or snow covering), (2) poor quality lane markings leading to insufficient contrast between the lane marking and the road, and/or (3) inconsistent lane markings which confuse the MV system (e.g., merge lanes, intersections).

3.2.2 Steering Wheel Inputs

Steering wheel inputs can be an indicator of drowsy driver impairment. Alert drivers will respond to lane deviations early with many small-amplitude steering movements which correct the car’s trajectory, whereas drowsy drivers will respond slowly to lane deviations and make large steering wheel movements in order to correct for larger lane deviations (Thiffault and Bergeron 2003). Most researchers agree that with an increase in fatigue-related impairment, a subsequent increase in the variability of steering control occurs. Beyond this general agreement there is little consensus as to what exact measure of steering is correlated with driver drowsiness (Fairclough 1997).

System developers must consider the possible confounds (e.g., road type, road crown, crosswinds, vehicle steering characteristics) when considering steering as an indicator of fatigue. While lane position is a product of steering input, it is the driver’s task to keep the

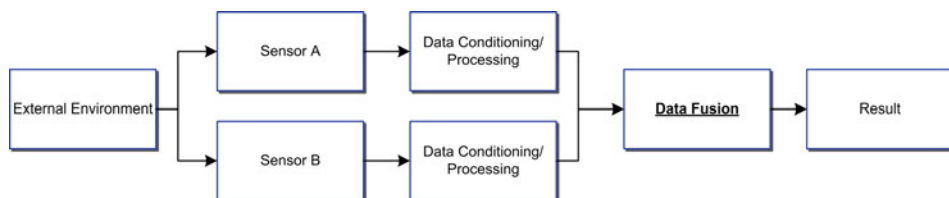
vehicle in the lane regardless of these confounds. For instance, an alert driver guiding a vehicle along a roadway may have to make frequent and high magnitude steering inputs to counter a crosswind, but the vehicle's change in lane position may not be indicative of these frequent steering corrections. However, as a driver becomes drowsy, lane-keeping behavior is quite likely to change because the driver is not sufficiently alert to counter the crosswinds. Consequently, lane-keeping behavior appears to be less sensitive to confounding factors while remaining sensitive to drowsiness.

3.3 Measures of Driver Drowsiness Summary

Ideally, DDASs will constantly monitor the driver and alert him or her when a safety threshold for fatigued/drowsy driving has been passed. However, detecting fatigue in drivers is difficult for a number of reasons. Methods such as assessing performance based on simple reaction-time tasks have not proven to be sufficient as a sole metric of fatigue and drowsiness (Baulk et al. 2008). Physiological measures such as EEG provide a much more reliable metric (Brookhuis and de Waard 2010; Kee et al. 2010). However, they are often difficult to implement in a laboratory setting or in a fielded DDAS. While physiological/ocular measures and performance data are reliable for identifying driver drowsiness with minimal false alerts (Misener et al. 2008; Liu et al. 2009), detecting drowsiness through eye data alone may be difficult as visual fatigue and general fatigue are often complicated to distinguish (Sullivan 2008). Finally, individual differences are present in nearly all of these drowsiness measures. It is important to understand the driver's *normal* pattern of behavior, whether the measure is eye opening size, blink characteristics, or steering style. A single set of drowsiness parameters for drowsiness may not be appropriate for all drivers.

4 Development of a Robust Drowsy Driver Assistance System

Fusion of data derived from driver-based and vehicle-based approaches may be more successful at reliably detecting fatigue (Morad et al. 2009). As previously mentioned, a single-measure method alone has resulted in issues related to reliability and/or problems with successful on-road implementation. In the previous section, many potential drowsiness indicators were identified and categorized as either driver-based or vehicle-based. Data fusion is the integration of information from different sources (e.g., sensors, vehicle network) into an output that has potentially greater value than the original data from the individual sources. ● Figure 34.1 shows a general diagram representing a simple data fusion work flow for reference. As shown in the figure, two or more sensors measure some aspect of the environment, after which the data may then separately be processed prior to being fused based on selected parameters. The final result is more reliable as it is based on the integration of information from two sources in contrast to just one source.



■ Fig. 34.1

Example of a simple data fusion work flow

The selection of multiple measures for successful data fusion first requires the identification of relevant functional specifications. A recommended set of functional specifications for the design of a DDAS is provided below.

- Accuracy** – First and foremost, the DDAS must accurately measure what it is intended to measure. This corresponds to the scientific principle of *validity*. Further, the DDAS should have an acceptable level of sensitivity to the variables it is measuring. The *sensitivity* of the system refers to the likelihood that the effect of a variable will be detected when that variable does, indeed, have an effect (Shaughnessy and Zechmeister 1994). Sensitivity is increased to the extent that error variation (e.g., false positives/negatives) is reduced. A driver may accept a small percentage of errors, but too many errors will ultimately reduce the driver's trust of the system (Lees and Lee 2007). This distrust can lead to annoyance and disregard of the system, which may ultimately lead to deactivation.
- Ability to Account for Common Driving Behaviors** – This criterion is closely related to the validity of the DDAS. Reasonable operator behaviors must be accounted for in the design of the system. For example, it must accommodate the need for drivers to check mirrors, look over one's shoulder, shift gears, make seat adjustments, reach for items within the cab, move into common driving postures (e.g., slouching, erect), etc. This is especially important for DDASs which include an eye-related measure. For instance, some currently available DDAS models will assume a driver's eyes are closed if the eyes cannot be detected, such as when the driver turns to check the vehicle's mirrors. Of course, this would be inaccurate if the driver's eyes are open when checking the mirrors.
- Reliability** – The DDAS must be reliable in that it *consistently* measures what it is intended to measure. Furthermore, the hardware and software of the system should require minimal maintenance. If the DDAS is often inconsistent and requires regular calibration or maintenance (beyond expectations), drivers may be less likely to trust, and therefore less likely to use, the system.
- Adaptability to Various Environmental Conditions** – An ideal DDAS must operate correctly in a variety of environmental conditions, both internal to the cab (e.g., temperature, ambient lighting, high noise levels, etc.) and external to the cab (e.g., time of day, adverse weather conditions, changing overhead lighting luminosities, dense or

sparse ambient traffic, varying highway geometry). This corresponds to the scientific principle of *generalizability*.

- *Adaptability to Various Driver Physical Characteristics* – Design for a wide variety of operator physical characteristics is critical for the system to be successful. The system should accommodate the widest range of physical characteristics of the operator, including demographic features (e.g., age and gender); physical features (e.g., face size, body height and size, eye color, skin color, facial features); transient features (e.g., if the driver wears a hat); and visual needs (e.g., corrected or uncorrected visual impairments, use of eyeglasses, sunglasses, contact lenses). This criterion also corresponds to the scientific principle of *generalizability*.
- *Ability to Detect a Change in the Vehicle Operator* – Somewhat related to the above is the system's ability to detect and account for multiple drivers, as most vehicles, for example commercial and light motor vehicles will frequently be operated by several individuals. The system must have the capability to self-calibrate and configure as needed to allow for sharing of the vehicle equipped with the DDAS. Hence, the system will ideally have the capability to identify a change in operator and adjust accordingly to detect changes in operators' physical characteristics.
- *Ability to Meet Fundamental Human Interface Needs* – To be effective, the system's information and, if appropriate, warning must be noticed, heard, understood, and accepted by the driver (Wickens et al. 1998) creating a user-centered approach to interface design. The system must support the driver, not hinder or create potentially hazardous situations. The information must maintain a balance of driver attention and resource allocation. Allocations among distraction and attention demands needed to safely operate the vehicle are critical. The success of an information/warning system will depend on the driver interface and how well the algorithm fits the driver's capabilities and preferences. The driver must understand that the information/warning is not intended to serve as an "alarm clock," but rather an indication of high likelihood of a hypo-vigilant state (a state of diminished arousal and decreased responsiveness to stimuli).
- *Non-Encumbering Design* – Related to the interface needs specified above, the DDAS must not obstruct the operator's field-of-view or access to necessary controls and displays required to operate the vehicle. Interaction with the system must be kept to a minimum, both for safety purposes and driver acceptance of the system. The system should not be overly distracting or require the driver to remove his/her hand(s) from the wheel for an extended period of time. In terms of driver acceptance, it seems to be evident that the greater the encumbrance of the system to the driver, the less likely the driver will be to accept the system. Ideally, such a system would require only limited, if any, intervention by the driver.
- *Need for Minimal Calibration* – While calibration based upon individuals' needs may be inevitable, ongoing calibration should be kept to a minimum. Closely related to non-encumbering features, any additional intervention required may lead to lack of acceptance. The calibration process should be simple and quick to implement as drivers may be less likely to deal with system calibration complexities.

- *Ability to Gather Data Continuously in Real time* – For the DDAS to serve its intended purpose, the system must operate in real time, thus having an acceptably short delay in updating status information and issuing warnings. Appreciable delays of the system will result in the reduction of the intended protection afforded by the system. Furthermore, it is clear that the system needs to have the ability to monitor driving continuously without major interruptions. Since drowsy driving can occur at any point during a drive, the system must be able to monitor the driver's condition throughout the entire driving session; hopefully, for the majority of time that the vehicle is in motion.
- *Cost-Effectiveness* – Once the system is well-defined and found to adequately meet all design and functional specifications, it should be made economically viable. It is generally the case that costs can be reduced once a viable prototype is developed. Ultimately, the benefits of the system must outweigh its costs, or the system will be dismissed by its potential consumers.

To summarize, the functional specifications described above are important to keep the driver safe and obtain/maintain user acceptance. These specifications may not be exhaustive, but it is the opinion of these authors that the functional specifications above are crucial for a reliable DDAS, and may also be applicable as a starting point for the design of other in-vehicle technologies.

4.1 Case Study of Prototype DDAS Utilizing a Data Fusion Approach

With these general design and functional specifications in place, the next general step in the development of a DDAS is to select the drowsiness-related indicators that are most applicable. As a case study, the current authors will describe an approach used to develop a prototype DDAS that fused data from two selected indicators of drowsiness (Hanowski et al. 2008b).

Past and present DDAS models have relied primarily on their ability to detect a single driver behavioral characteristic such as ocular movements or steering, or vehicle-based measures such as lane position or line crossing. For example, one driver-based measure used in some drowsiness monitors is slow eye closure (i.e., PERCLOS). However, tests with some specific technologies have found periodic intervals of data loss in various conditions including when: the driver has eyewear (e.g., glasses, sunglasses), the driver's eyes are not within the system's field-of-regard because of normal visual scanning patterns (e.g., mirror checks), or the driver is performing a secondary task (e.g., looking down at the speedometer) that can be misread by the system as an eye closure (Wierwille et al. 2003). Lane position, like slow eye closures, is another metric that has been used to identify drowsy driving (operator/vehicle performance parameter). And, like slow eye closures, is not a completely robust measure in the real-world. For example, lane position data loss can occur when the system fails to read the lane's edge markings due to

(1) weather, such as rain or snow covering, (2) poor quality lane markings leading to insufficient contrast between the lane marking and the road, and/or (3) inconsistent lane markings, such as merge lanes or intersections, which may confuse the sensors. Depending on how data loss is handled by the system, the result can be a missed occurrence of drowsy driving or false alarms that indicate that the driver is drowsy when actually alert. Handling false alarms can also be a difficult problem as too many false alarms can diminish the users trust, confidence, and acceptance of the technology (Bliss and Acton 2003; Lees and Lee 2007).

4.1.1 Drowsiness Indicator Selection


The objective of this example prototype DDAS development effort was to develop a robust system that combined drowsiness metrics to alleviate some of the issues associated with a single-measure device. Based on an information gathering task, it was determined that a machine-vision (MV)-based eye monitoring technology in combination with analysis of operator/vehicle performance parameters would provide a strong combined metric to reliably assess driver drowsiness. As noted, the problem with single-measure drowsiness detection systems is that when there is a loss of data, the system becomes unreliable. For example, single-measure systems that are based on eye closure metrics may have data loss if the driver is wearing eyeglasses or there is high ambient illumination (Hanowski et al. 2008a). However, a multi-measure approach, that uses multiple distinct sensors, can provide not only a backup system but also provide an integrated approach whereby both measures provide more robust drowsiness detection than either measure could alone.

Keeping with the previously recommended development method, a comprehensive information gathering effort was conducted to identify salient indicators of driver drowsiness. These included (1) driver-based measures of drivers' current state of well-being ascertained by the physiological changes in ocular movements, EEG, pupil occlusion, etc., (2) vehicle-based measures of driving performance using variables such as speed, lateral lane position, time-to-line crossing, lane drift, steering movements, etc., and (3) a combination of person- and vehicle-based measures.

Another key finding was that PERCLOS is, arguably, the most valid driver-based drowsiness measure that can be used in a real-world driving environment (Dinges and Grace 1998; Knipling 1998; Wierwille 1994). However, the sensor used to assess PERCLOS may not always provide a reliable measure and this has been an issue with PERCLOS monitors that may not work reliably with certain driver characteristics (e.g., glasses) or certain environments (e.g., high ambient illumination) (Lees and Lee 2007). As such, vehicle-based measures can be used as a secondary measure to PERCLOS that, when PERCLOS is being determined reliably, can provide a more enhanced drowsiness measure and, perhaps more importantly, when PERCLOS is not being determined reliably, can serve as a backup measure.

As mentioned, a scan of current vehicle-based drowsiness measures revealed the lack of a single, “best” measure. Vehicle-based measures that previous research has found promising included lane position/line crossing (Wierwille 1994; Chatterjee et al. 1996; O’Hanlon and Kelley 1977; Tijerina et al. 1999) and steering (Fairclough 1997). For the current effort, Lane Position was selected as the vehicle-based measure and has been examined together with PERCLOS in past studies (Wierwille 1994; Pilutti and Ulsoy 1997).

4.1.2 Fusion of Data from Two Drowsiness Indicators

There were two key tasks associated with this example DDAS data fusion effort: (1) model derivation and (2) sensor data integration. The model derivation involved developing an approach for combining the PERCLOS and Lane Position metrics, setting thresholds for alerting, and presentation of the alerts to the driver.  Figure 34.2 shows a diagram of the DDAS prototype data fusion work flow for reference. Lane Deviation (i.e., distance between the vehicle’s centerline and lane centerline) was used as the specific Lane Position measure. Although Lane Deviation was the measure used in this effort, other lane position measures (e.g. time out of lane, departure angle) may produce similar or better results with the appropriate implementation.

A straightforward mathematical model was derived to exercise the prototype DDAS. This preliminary algorithm used a 3×3 lookup table to provide an associative array of drowsiness categories that correspond to the specified levels of PERCLOS and Lane Offset (GREEN on the Driver/Vehicle Interface [DVI], represents a rating of 1, YELLOW represents a rating of 2, and RED represents a rating of 3). The algorithm is mathematically defined as:

$$X = (\text{DDMS PERCLOS_Category}) + \ln(\text{LaneDeviation_Category}) \quad (34.1)$$

Where X is classified as one of the following drowsiness categories:

$X < 2$ = Drowsiness Category 1 (depicted as GREEN on the DVI)

$2 \leq X < 3$ = Drowsiness Category 2 (depicted as YELLOW on the DVI)

$X \geq 3$ = Drowsiness Category 3 (depicted as RED on the DVI)

By taking the natural logarithm of the Lane Deviation value, the algorithm is, in effect, exponentially increasing the power of the PERCLOS metric as the integrated Drowsiness value increases. The first value in the equation is associated with the PERCLOS category and the second value is associated with the Lane Deviation category. Based on these calculations, the DDAS output to the DVI is depicted by the GREEN, YELLOW, and RED colors.

The specific levels of the threshold criteria used to establish the categories of PERCLOS and Lane Deviation are intended to elicit a specific response from the drowsiness monitor. While these threshold criteria are founded on previous research (Wierwille et al. 2003) and expert judgment, they should be considered preliminary.

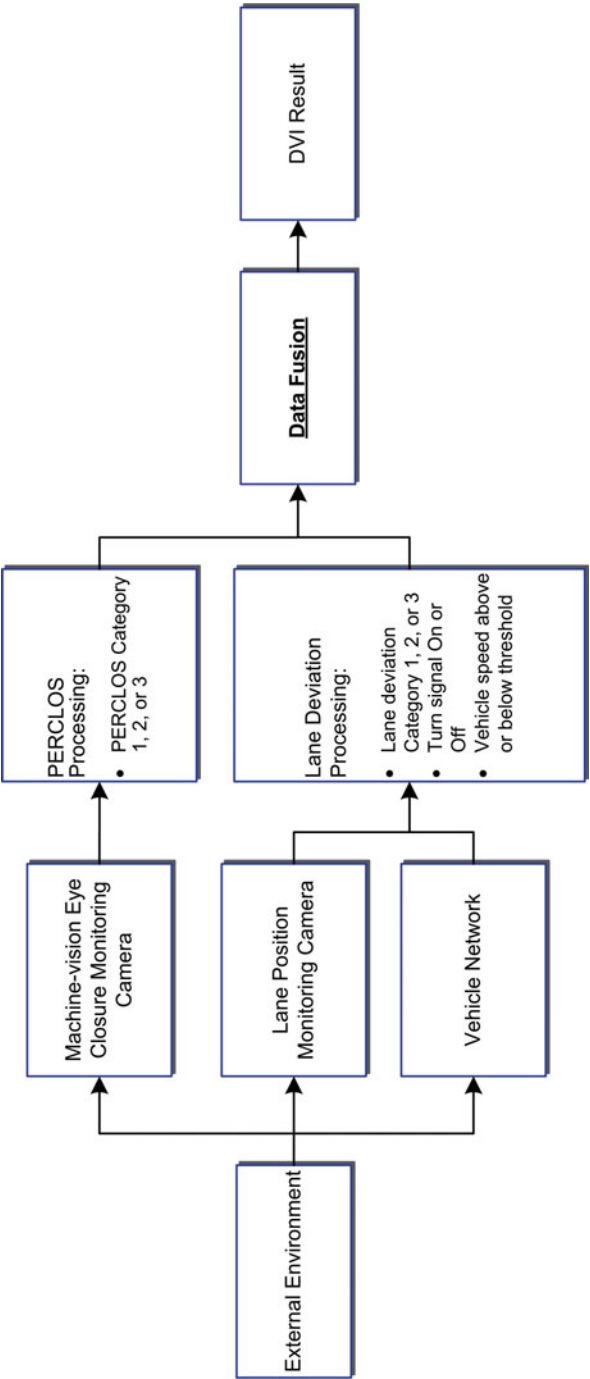


Fig. 34.2 DDAS prototype data fusion work flow

4.1.3 Testing and Evaluation of Prototype DDAS

The operational performance of the prototype DDAS was assessed during a dynamic on-road evaluation under varying conditions of ambient illumination (i.e., day and night with and without overhead street lighting), eyewear (i.e., prescription glasses and sunglasses), and skin complexion (i.e., light and dark). The MV slow eye closure and MV lane position sensors were integrated and installed in a Class-8 tractor and tested on a closed-test track. The purpose of the testing was to “exercise” the system and assess the performance of the DDAS in terms of ideal functional specifications (e.g., accuracy and sensitivity) to determine the effectiveness of the multiple sensor integration approach. It is important to note that the on-road evaluation was not to test human behavior and performance, but rather, the evaluation focused on exercising the prototype system to determine its operational envelope (i.e., those operational conditions in which the system does and does not work effectively).

Because the prototype system used a combination of eye closure and lane position sensors, the evaluation approach involved separating the driving task from the eye closure tasks. The driver for all test sessions was a research technician with a CDL. He was trained to perform consistent lane deviation maneuvers according to the test protocol, which assessed the effectiveness of the MV lane position sensor. At the same time, nine other participants with varying physical facial features (e.g., skin complexion) and eyewear sat in the passenger seat and performed prescribed eye closures to exercise the MV slow eye-closure sensor. Additional participant tasks were set up to evaluate the system for driver behaviors that may elicit false alarms (e.g., mirror checks). This two-participant approach served as an effective and efficient method that allowed for multiple stimuli to be presented to the system simultaneously in a manner that was repeatable and safe for the research participants.

The key finding was that a multiple sensor data fusion approach provides a more robust method for assessing driver drowsiness. ♦ [Table 34.1](#) summarizes the operational performance of the DDAS against the functional specifications outlined earlier. For each functional specification attribute, an indicator is provided to show that the performance for that prototype DDAS component was: (1) at an acceptable level of accuracy, (2) at a marginal level of accuracy, or (3) at an unacceptable level of accuracy. The final column in ♦ [Table 34.1](#) provides an indication of whether the integrated prototype DDAS performed to a higher level than the two independent sensors alone. Neither the slow eye-closure sensor nor the lane position sensor performed perfectly under this study’s real-world conditions; however, as an integrated system, the DDAS performed successfully for 15 of the 17 functional specifications tested. The primary limitations of the MV slow eye-closure sensor were: (1) eyewear, (2) nonuniform illumination on the face, and (3) head/body motions caused by vehicle ride motions and driving tasks. The primary limitation of the MV lane position sensor was instances of low contrast ratio between the lane’s markings and the surrounding scene. With further improvements, the integrated DDAS could expand its capabilities into the two areas of marginal performance. It is believed that with further system improvements, an integrated DDAS would be effective across all necessary functional specifications.

■ Table 34.1

Summary of DDAS operational performance

Functional specifications	Description	DDAS PERCLOS estimate	DDAS lane deviation estimate	Fused data DDAS performance
Accounts for common driving behaviors	Visual scanning	Marginal	Acceptable	Successful
	Shifting gears	Marginal	Acceptable	Successful
	Adjusting posture	Unacceptable	Acceptable	Successful
	Reaching for Items	Unacceptable	Acceptable	Successful
	Common driving postures	Marginal	Acceptable	Successful
Environmental conditions	Daytime illumination (≥ 753.2 lux)	Acceptable	Acceptable	Successful
	Nighttime illumination (≤ 0.5 lux)	Marginal	Marginal	Improvement area
	Nighttime illumination with artificial overhead lighting (≥ 0.6 lux and < 753.2 lux)	Marginal	Marginal	Improvement area
Accounts for various driver physical characteristics	No eyewear	Acceptable	Acceptable	Successful
	Prescription eyeglasses	Unacceptable	Acceptable	Successful
	Sunglasses	Unacceptable	Acceptable	Successful
	Light skin complexion	Acceptable	Acceptable	Successful
	Dark skin complexion	Acceptable	Acceptable	Successful
Accounts for multiple drivers		Acceptable	Acceptable	Successful
Non-encumbering design	Not obstruct operator's field of view	Acceptable	Acceptable	Successful
	Access to necessary controls	Not assessed	Not assessed	Not assessed
Minimal calibration	Minimum Ongoing Calibration	Acceptable	Acceptable	Successful
Real-time data gathering	Acceptable short delays in updating status and issuing warnings.	Acceptable	Acceptable	Successful

5 Conclusion

Currently, no single measure reliably predicts driver drowsiness 100% of the time. Therefore, a multi-measure approach will provide a more reliable detection of the onset of driver drowsiness. The general idea behind a data fusion approach to DDAS development was that multiple measures of drowsiness could be combined to create a measure that would be more robust, in a real operating environment, than a single

drowsiness measure. The example prototype DDAS development effort was provided to demonstrate that a working prototype, that combined two drowsiness measures, could be developed and tested on a real road. To this end, the combined drowsiness measure for the current example was slow eye closure + lane position.

This work has been supported by recent commercial efforts that are transiting this notion of data fusion from research to practice. One major automotive company has begun to offer driver assistance systems that use multiple measures of drowsiness. Mercedes released its “Attention Assist” system in 2009. The Attention Assist is designed to detect the initial signs of tiredness by comparing real-time driving patterns with a baseline driver profile. The driver warning is dependent on different measures such as linear and lateral acceleration, pedal operation, vehicle speed, time of day, duration of the trip, and the steering response (Zer Customs 2007).

As more is learned about the factors that lead to the onset of driver drowsiness, drowsiness detection system developers can continue to strengthen their prediction models through a multi-measure, fused data approach for detecting driver drowsiness. For example, taking into consideration the amount of sleep the driver had the night before, time since last sleep, and circadian low periods are just a few additional measures that could be added to enhance driver drowsiness prediction models. The combination of measures which best predicts or indicates the onset of driver drowsiness and the corresponding impairment remains to be answered.

References

- Åkerstedt T, Peters B, Anund A, Kecklund G (2005) Impaired alertness and performance driving home from the night shift: a driving simulator study. *J Sleep Res* 14(1):17–20
- Artaud T, Planque S, Lavergne C, Cara H, de Lepine P, Tarrière C, Gueguen BM (1994) An on-board system for detecting lapses of alertness in car driving. In: *Proceedings of the 14th international conference on enhanced safety of vehicles 1*, Munich
- Baulk SD, Biggs SN, Reid KJ, van den Heuvel CJ, Dawson D (2008) Chasing the silver bullet: measuring driver fatigue using simple and complex tasks. *Accid Anal Prev* 40:396–402
- Bhuiyan M (2009) Driver assistance systems to rate drowsiness: a preliminary study. In: Nakamatsu K, Phillips-Wren G, Jain L, Howlett R (eds) *New advances in intelligent decision technologies*, vol 199. Springer, Berlin/Heidelberg, pp 415–425
- Bliss JP, Acton SA (2003) Alarm mistrust in automobiles: how collision alarm reliability affects driving. *Appl Ergon* 34(6):499–509
- Brookhuis K (1995) Driver impairment monitoring by physiological measures. In: Hartley L (ed) *Fatigue and driving: driver impairment, driver fatigue and driving simulation*. Taylor and Francis, London, pp 181–188
- Brookhuis KA, de Waard D (2010) Monitoring drivers’ mental workload in driving simulators using physiological measures. *Accid Anal Prev* 42(3):898–903. doi:10.1016/j.aap.2009.06.001
- Brookhuis KA, Louwerens JW, O’Hanlon JF (1986) EEG energy-density spectra and driving performance under the influence of some anti-depressant drugs. In: O’Hanlon JF, de Gier JJ (eds) *Drugs and driving*. Taylor and Francis, London, pp 213–221
- Caffier PP, Erdmann U, Ullsperger P (2003) Experimental evaluation of eye-blink parameters as a drowsiness measure. *Eur J Appl Physiol* 89:319–325
- Chatterjee A, Cadotte E, Stamatiadis N, Sink H, Venigalla M, Gaides G (1994) Driver-related factors involved with truck accidents. *Southeastern*

- Transportation Center Project No. 23385-019. Southeastern Transportation Center, Knoxville
- Chatterjee A, Cadotte E, Stamatiadis N, Sink H, Venigalla M, Gaides G (1996) Driver-related factors involved with truck accidents. *J Saf Res* 27(1):56
- Dinges DF (1995) An overview of sleepiness and accidents. *J Sleep Res* 4:4–14
- Dinges DE, Grace R (1998) PERCLOS: a valid psychophysiological measure of alertness as assessed by psychomotor vigilance. Report No. FHWA-MCRT-98-006. US Department of Transportation, Washington, DC
- Erwin CW (1976) Studies of drowsiness: final report. The National Driving Center, Durham
- Erwin CW, Volow MR, Gray B (1973) Psychophysiological indices of drowsiness. SAE, New York
- Fairclough SH (1997) Monitoring driver fatigue via driving performance. In: Ian Noy (ed) *Ergonomics and safety of intelligent driver interfaces*. Lawrence Erlbaum Assoc, Hillsdale, pp 363–379
- Folkard S (1997) Black times: temporal determinants of transport safety. *Accid Anal Prev* 29(4):417–430
- Hanowski RJ, Blanco M, Nakata A, Hickman JS, Schaudt WA, Fumero MC, Olson RL, Jermeland J, Greening M, Holbrook GT, Knipling RR, Madison P (2008) The drowsy driver warning system field operational test, data collection. Contract DTNH22-00-C-07007, Task Order 14, final report. National Highway Traffic Safety Admin., Washington, DC
- Hanowski RJ, Bowman DS, Wierwille WW, Alden A, Carroll R (2008) PERCLOS+: development of a robust field measure of driver drowsiness. In: *Proceedings of the 15th world congress on intelligent transport systems* (CD-ROM)
- Horne JA, Reyner LA (1995) Sleep related vehicle accidents. *Br Med J* 310:565–567
- Ingre M, Åkerstedt T, Peters B, Anund A, Kecklund G (2006) Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *J Sleep Res* 15:47–53
- Kecklund GA, Åkerstedt T (1993) Sleepiness in long distance truck driving: ambulatory EEG study of night driving. *Ergonomics* 36:1007–1017
- Kee S, Tamrin SBM, Goh Y (2010) Driving fatigue and performance among occupational drivers in simulated prolonged driving. *Glob J Health Sci* 2(1):167–177
- Knipling R (1998) The technologies, economics, and psychology of commercial motor vehicle driver fatigue management. Presented at the 1998 ITS America's eighth annual meeting and exposition, Detroit
- Knipling RR, Wierwille WW (1994) Vehicle-based drowsy driver detection: current status and future prospects. In: *Proceedings of the IVHS America 1994 annual meeting*, Atlanta, pp 245–256
- Lal SKL, Craig A, Boord P, Kirkup L, Nguyen H (2003) Development of an algorithm for an EEG-based driver fatigue countermeasure. *J Saf Res* 34(3):321–328
- Lees MN, Lee JD (2007) The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics* 50(8):1264–1286
- Lenne MG, Triggs TJ, Redman JR (1997) Time of day variations in driving performance. *Accid Anal Prev* 29:431–437
- Liu CC, Hosking SG, Lenné MG (2009) Predicting driver drowsiness using vehicle measures: recent insights and future challenges. *J Saf Res* 40:239–245
- Lyznicki JM, Doege TC, Davis RM, Williams MA (1998) Sleepiness, driving, and motor vehicle crashes. *J Am Med Assoc* 279(23):1908–1913
- Misener J, Nowakowski C, O'Connell J, Murray J (2008) Onboard monitoring and reporting for commercial vehicle safety (OBMS) phase II: field operational test. Report No. UCB-ITS-PRR-2008-18. California PATH, Berkeley
- Morad Y, Barkana Y, Zadok D, Hartstein M, Pras E, Bar-Dayana Y (2009) Ocular parameters as an objective tool for the assessment of truck drivers fatigue. *Accid Anal Prev* 41:856–860
- National Transportation Safety Board (1990) *Fatigue, alcohol, other drugs, and medical factors in fatal-to-the-driver heavy truck crashes*, vol 1. NTSB publications. <http://www.nts.gov/publictn/1990/ss9001.htm>. Accessed 6 June 2011
- O'Hanlon JF, Kelley GR (1977) Comparison of performance and physiological changes between drivers who perform well and poorly during prolonged vehicular operation. In: Mackie RR (ed) *Vigilance*. Plenum, New York, pp 189–202
- O'Hanlon JF, Kelly GR (1974) A psycho-physiological evaluation of devices for preventing lane drift and run-off-road accidents, Technical report 1736-F. Human Factors Research Inc., Santa Barbara Research Park, Goleta
- Picot A, Charbonnier S, Caplier A (2010) Drowsiness detection based on visual signs: blinking analysis

- based on high frame rate video. Paper presented at the 2010 IEEE instrumentation and measurement technology conference (I²MTC), Austin, 3–6 May 2010
- Pilutti T, Ulsoy G (1997) Identification of driver state for lane-keeping tasks: experimental results. In: Proc. of the American Control Conf, New Mexico, pp 3370–3374
- Shaughnessy JJ, Zechmeister EB (1994) Research methods in psychology. McGraw-Hill, New York
- Skipper JH, Wierwille WW (1985) An investigation of low-level stimulus-induced measures of driver drowsiness. In: Gale AG, Freeman MH, Haslegrave CM, Smith P, Taylor SP (eds) Proceedings of the conference on vision in vehicles. North Holland Elsevier Science, Nottingham, pp 139–148. (Discussion by Wierwille WW, 217–218)
- Skipper JH, Wierwille WW (1986) Drowsy driver detection using discriminant analysis. *Human Factors* 28(5):527–540
- Stern JA, Boyer D, Schroeder D (1994) Blink rate: a possible measure of fatigue. *Human Factors* 36(2):285–297
- Stutts JC, Wilkins JW, Vaughn BV (1999) Why do people have drowsy driving crashes? Input from drivers who just did. AAA Foundation for Traffic Safety, Falls Church
- Sullivan JM (2008) Visual fatigue and the driver, technical report. University of Michigan Transportation Research Institute, Ann Arbor
- Thiffault P, Bergeron J (2003) Monotony of road environment and drivers fatigue: a simulator study. *Accid Anal Prev* 35(3):381–391
- Tijerina L, Glecker M, Stoltzfus D, Johnston S, Goodman MJ, Wierwille WW (1999) A preliminary assessment of algorithms for drowsy and inattentive driver detection on the road. Presented at the 9th Intelligent Trans. So. of America, Washington, DC
- Torsvall L, Åkerstedt T (1987) Sleepiness on the job: continuously measured EEG changes in train drivers. *Electroencephalogr Clin Neurophysiol* 66:502–511
- Volow MR, Erwin CW (1973) The heart rate variability correlates of spontaneous drowsiness onset. International Automotive Engineering Congress, Detroit
- Wickens C, Gordon S, Liu Y (1998) An introduction to human factors engineering. Addison-Wesley Longman, New York
- Wierwille WW (1994) Overview of research on driver drowsiness definition and driver drowsiness detection. In: Proc. of the 14th Int. Conf. on enhanced safety of vehicles, Munich, pp 462–468
- Wierwille WW, Hanowski RJ, Olson RL, Dinges DF, Price NJ, Maislin G, Powell IV JW, Ecker AJ, Mallis MM, Szuba MP, Ayoob A, Grace R, Steinfeld A (2003) NHTSA drowsy driver detection and interface project. Contract DTNH22-D-00-07007, Task Order 7, Final Rep. National Highway Traffic Safety Administration, Washington, DC
- Williamson A, Feyer A, Friswell R (1996) The impact of work practices on fatigue in long distance truck drivers. *Accid Anal Prev* 28(6):709–719
- Wylie CD, Schultz T, Miller JC, Mitler MM, Mackie RR (1996) Commercial motor vehicle driver fatigue and alertness study: technical summary. Rep. No. FHWA-MC-97-001. US Department of Transportation, Federal Highway Administration, Washington, DC
- Zer Customs (2007) Mercedes attention assist. Retrieved from <http://www.zercustoms.com/news/Mercedes-Attention-Assist.html>. Accessed 3 June 2011

35 Drowsy Driver Posture, Facial, and Eye Monitoring Methods

Jixu Chen¹ · Qiang Ji²

¹Visualization and Computer Vision Lab, GE Global Research Center, Niskayuna, NY, USA

²Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

1	<i>Introduction</i>	915
2	<i>Background</i>	918
2.1	Facial Activity Recognition	918
2.1.1	Facial Feature Point Tracking	919
2.1.2	Facial Expression Recognition	919
2.2	Eye Gaze Estimation	920
3	<i>Facial Feature Point Tracking</i>	921
3.1	Facial Feature Point Representation	921
3.2	Pose-Dependent Active Shape Model	923
3.2.1	Active Shape Model	923
3.2.2	Robust Face Pose Estimation	924
3.2.3	Face Shape Compensation	925
4	<i>A Unified Framework for Simultaneous Facial Activity Tracking</i>	926
4.1	DBN Model Parameterization	927
4.1.1	Expression Level	927
4.1.2	AU Level	928
4.1.3	Facial Feature Level	929
4.2	DBN Inference: Simultaneous Facial Feature Tracking and Expression Recognition	931
4.3	Result on the Cohn-Kanade Database	931
5	<i>3D Gaze Estimation Based on Facial Feature Points Without IR Illumination</i>	932
5.1	Step 1. Facial Feature Point Tracking and Face Pose Estimation	933
5.2	Step 2. Estimate the 3D Model Point M	934

5.3 Step 3. Compute 3D Position of C 935

5.4 Step 4. Compute P 935

5.5 Step 5. Compute Gaze 935

5.6 Gaze Estimation Result 936

6 Conclusion 937

Abstract: This chapter presents a real-time computer vision system for monitoring drowsy driver. It uses one remotely located charge coupled device (CCD) camera to acquire video of the driver's face. From the video, various computer vision algorithms are employed to simultaneously, nonintrusively, and in real time recognize the facial behaviors that closely relate to the driver's level of vigilance. The facial behaviors include rigid head movement (characterized by 3D face pose), nonrigid facial muscular movement (characterized by facial expressions), and eye gaze movement. The system was tested in a simulating environment with different subjects and it was found robust, reliable, and accurate in characterizing facial behaviors.

1 Introduction

The ever-increasing number of traffic accidents in the USA due to a diminished driver's vigilance level has become a problem of serious concern to society. Drivers with a diminished vigilance level suffer from a marked decline in their abilities of perception, recognition, and vehicle control, and therefore pose serious danger to their own life and the lives of other people. Statistics show that a leading cause for fatal or injury-causing traffic accidents is due to drivers with a diminished vigilance level. In the trucking industry, 57% fatal truck accidents are due to driver fatigue. It is the number one cause for heavy truck crashes. Seventy percent of American drivers report driving fatigued. With the ever-growing traffic conditions, this problem will further deteriorate. For this reason, developing systems actively monitoring a driver's level of vigilance and alerting the driver of any insecure driving conditions is essential to prevent accidents.

Many efforts (Ishii et al. 1987; Saito 1992; Yammamoto and Higuchi 1992; Saito et al. 1994; Boverie et al. 1998; Anon 1999; Smith et al. 2000; Ji et al. 2004, 2006) have been reported in the literature for developing active safety systems intended for reducing the number of automobile accidents due to reduced vigilance. Among different techniques, the best detection accuracy is achieved with techniques that measure physiological conditions like brain waves, heart rate, and pulse rate (Saito 1992; Yammamoto and Higuchi 1992). Requiring physical contact with drivers (e.g., attaching electrodes) to perform these techniques are intrusive, causing annoyance to drivers. Good results have also been reported with techniques that monitor eyelid movement and gaze with a head-mounted eye tracker or special contact lens. Results from monitoring head movement (Saito 1992) with a head-mount device are also encouraging. These techniques, though less intrusive, are still not practically acceptable.

People in fatigue exhibit certain visual behaviors easily observable from changes in their facial features like the eyes, head, and face. Typical visual characteristics observable from the image of a person with reduced alertness level include slow eyelid movement (Dinges et al. 1998), smaller degree of eye openness (or even closed), frequent nodding (Anon 1998), yawning, gaze (narrowness in the line of sight), sluggishness in facial expression, and sagging posture. To make use of these visual cues, another increasingly popular and noninvasive approach for monitoring fatigue is to assess a driver's vigilance


level through visual observation of his/her physical conditions using a camera and state-of-the-art technologies in computer vision.

Many real-time video-based nonintrusive fatigue monitoring systems have been proposed. For example, Ishii et al. (1987) introduced a system for characterizing a driver's mental state from his facial expression. Saito et al. (1994) proposed a vision system to detect a driver's physical and mental conditions from line of sight (gaze). Boverie et al. (1998) described a system for monitoring driving vigilance by studying the eyelid movement. Their preliminary evaluation revealed promising results of their system for characterizing a driver's vigilance level using eyelid movement.

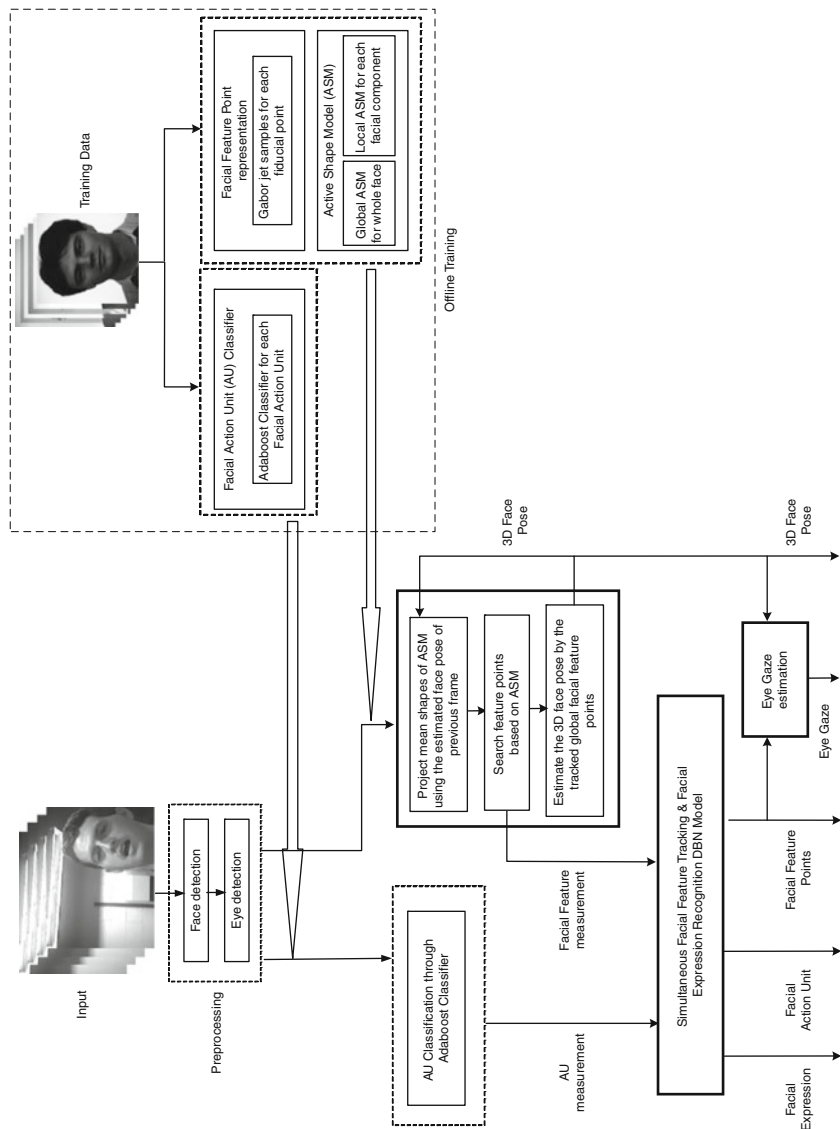
Ji et al. (2004, 2006) have developed a prototype computer vision system for monitoring driver vigilance. The main components of the system include remotely located charge coupled device (CCD) video cameras and several infrared (IR) illuminators. Various computer vision algorithms are proposed for simultaneous, real-time and nonintrusive monitoring of visual biobehaviors that typically characterize a driver's level of vigilance and attention. The parameters computed from these visual cues are subsequently combined probabilistically using a Bayesian Network to form a composite index that can robustly, accurately, and consistently characterize a driver's alertness level. For implementation of their system in a vehicle, they propose to use two CCD cameras embedded on the dashboard. The first camera is a narrow-angle camera focused on the driver's eyes to monitor eyelid and gaze movements, and the second camera is a wide-angle camera aimed at the driver's head to track and monitor head movement and facial expression.

Despite the success of the existing video-based systems for drowsy driver monitoring, accurate, robust, and efficient recognition of facial behaviors remains challenging. For facial expression recognition, the challenge arises from large facial deformations due to pose variation or facial expression change. For eye gaze tracking, most current gaze tracking systems need to use the IR lights to illuminate the face region and build the corneal reflection (glints) on the eye corneal surface. These gaze tracking systems based on IRs, however, can be affected by the sunshine in real-world applications and typically do not work outdoors.

In this chapter, we propose the computer vision algorithms to perform the facial feature points tracking, face pose estimation, facial expression recognition, and eye gaze tracking simultaneously from one camera without IR lights.

The flowchart in  Fig. 35.1 demonstrates our system. The modules in thick solid rectangles are the three main components of the system:

1. First, we propose a facial feature tracking algorithm to track 28 fiducial points surrounding facial components: eyebrows, eyes, nose, and mouth. This tracker provides the basic measurement of facial movement, which is used in the following face pose estimation, expression recognition, and eye gaze estimation.
2. Second, in order to recognize facial expression and to improve the accuracy of facial feature tracking, we propose a simultaneous facial activity tracking framework. Here, we model the facial activity in the following three levels (from local to global): In the



■ Fig. 35.1

The flowchart of the automatic facial feature tracking system based on the multistate hierarchical shape model

bottom level, facial feature point tracking focuses on the prominent facial points; in the middle level, local facial action units (AUs) characterize the specific behaviors of local facial muscular movements (e.g., mouth open and eyebrow raised); in the top level, global facial expression (e.g., surprise and happiness) represents the global facial movement. Different from previous methods which track the facial activities in three levels separately, we propose to model their relationships as well as their interplay using a Dynamic Bayesian network (DBN), and track them simultaneously. The proposed framework can improve the tracking (or recognition) performance in all three levels.

3. Finally, based on the tracked facial feature points and face pose, we introduce an eye gaze estimation algorithm which can estimate the person's line-of-sight without using IR lights.

The outputs of the proposed computer vision system include: global facial expression, local facial action unit (AU), facial feature point movement, face pose, and eye gaze movement. Based on the previous study (Ji et al. 2006), these human behaviors are closely related to a person's level of fatigue. For example, the facial AU27 (mouth stretch) can be used as a good indicator of yawning. The feature point movement on the eyelid can be used to extract the Percentage of Eye Closure Over Time (PERCLOS) and the Average Eye Closure Speed (AECS). PERCLOS has been validated (Dinges et al. 1998) and is found to be the most valid ocular parameter for characterizing fatigue. Face pose movement and eye gaze contains information about one's attentions. Certain face pose movements, such as head tilts, and narrow gaze are also good indicators of a fatigued driver.

The rest of the chapter is arranged as follows. ➤ Section 2 provides the background and the related works of facial feature point tracking, facial expression recognition, and eye gaze estimation. ➤ Section 3 presents the facial feature point tracking and face pose estimation algorithm. ➤ Section 4 introduces the framework to recognize facial expression and improve facial feature point tracking simultaneously. In ➤ Sect. 5, we propose a gaze estimation method based on the facial feature points without IR illumination. The chapter concludes in ➤ Sect. 6.

2 Background

2.1 Facial Activity Recognition

Facial activity, which is the major source of information for understanding one's intention and emotional state, has drawn growing attention in industry and academia. In recent years, plenty of the computer vision techniques have been developed to track or recognize the facial activities in three levels. In the bottom level, we can capture a detailed face shape by tracking the facial feature points, which are the prominent landmarks surrounding facial components. But sometimes we are only interested in the higher level information, such as some meaningful facial behaviors, for example, mouth open, eye close, eyebrow raiser, etc. Based on the psychological studies of Ekman's facial action coding system (FACS)

(Ekman and Friesen 1978), we can use the facial action units (AUs) resulting from the local facial muscular movements to characterize these facial behaviors. In the top level, facial expression analysis attempts to recognize six basic facial expressions, that is, happiness, surprise, sadness, fear, disgust, and anger (Ekman and Friesen 1978). These basic expressions represent the global facial muscular movement.

2.1.1 Facial Feature Point Tracking

Accurate localization and tracking local facial feature points is important to subsequent facial behavior analysis and recognition. In general, the facial feature point tracking techniques can be classified into model-free and model-based algorithms. The model-free algorithms only use the general purpose point trackers (Tomasi and Kanade 1991; Bourel et al. 2000). However, the point trackers are susceptible to the inevitable errors due to noise or occlusion. Recently, extensive work has been focused on the model-based facial feature point tracking which utilizes the facial shape constraints, such as active shape model (ASM) (Cootes et al. 1995), active appearance model (AAM) (Cootes et al. 2001), and elastic bunch graph matching (EBGM) (Wiskott et al. 1997).

Research has shown that these model-based methods are effective when tracking nearly frontal face. However, in real-world applications, facial tracking is still challenging due to large pose variation and significant facial expression change. To model the nonlinear deformations of face shape caused by pose/expression change, we propose two algorithms: (1) To resolve face pose changes, we propose to first estimate the face pose and then use the estimated pose to correct shape model in every frame (➊ Sect. 3); (2) To resolve the facial expression changes, we propose a unified framework to track facial activities in different levels simultaneously, that is, perform facial feature point tracking, AU recognition and expression simultaneously (➋ Sect. 4).

2.1.2 Facial Expression Recognition

In the literature, the facial expression recognition systems are usually focused on either the recognition of the six typical expressions (Cohen et al. 2003; Zhang and Ji 2005; Shan et al. 2006; Buenaposada et al. 2008; Dornaika and Davoine 2008) or the recognition of the AUs (Lien et al. 1999; Donato et al. 1999; Bazzo and Lamar 2004; Bartlett et al. 2005, 2006; Tong et al. 2007a, b, 2010). Since the temporal facial evolvement brings more information of the expression/AUs, most attention has been given to the temporal approach which tries to recognize the expression/AUs in the video. In general, an expression recognition system consists of two key stages: First, various facial features are extracted to represent the facial gestures or facial movements, for example, dense optical flow are used by Lien et al. (1999) and Donato et al. (1999) to detect the direction and magnitude of the facial movements; Bartlett et al. (2005) convolves the whole face image by a set of Gabor wavelet kernels, and the resulting Gabor wavelet magnitude are used as facial features.

Given the extracted facial features, the expression/AUs are identified by recognition engines, such as the Neutral Networks (Donato et al. 1999; Tian et al. 2001; Bazzo and Lamar 2004), Hidden Markov Models (Lien et al. 1999), Adaboost classifier (Bartlett et al. 2005, 2006) and Bayesian networks (Cohen et al. 2003; Zhang and Ji 2005; Tong et al. 2007a, b).

The facial feature point tracking, AU recognition and expression recognition represent the facial activity in three levels, and they are interdependent problems. For example, the facial feature point tracking can be used in the feature extraction stage in expression recognition, and the expression recognition results can provide a prior shape information in the model-based facial feature point tracking. However, most current systems only track the facial activities in one or two levels, and track them separately, ignoring their interactions. In addition, the computer vision measurements in each level are always uncertain and ambiguous. They are uncertain because of the presence of noise, occlusion, and of the imperfect nature of the vision algorithm. They are ambiguous because they only measure certain aspects of the visual activity, for example, the facial feature point tracking usually depends on local search and it is prone to drift, on the other hand, the expression recognition depends on global features but loses some details. Therefore, one expects to better infer the facial activities by systematically combining the measurements from multiple sources.

The idea of combining tracking with other problem has been attempted before, such as simultaneous face tracking and recognition (Zhou et al. 2003). Most recently, Fadi et al. (Dornaika and Davoine 2008) proposed a simultaneous facial action tracking and expression recognition algorithm. In their algorithm, they track the deformation of a 3D face mesh. By utilizing the dynamic of the expression and modeling the relationships between the expression and the 3D face mesh, the tracking performance is improved. However, since their model is subject-dependent, they need to train the model for each subject. Furthermore, they only model six basic expressions, which is a very small subset of human expressions.

We propose a unified probabilistic framework based on the dynamic Bayesian network (DBN) to explicitly model the relationships among the three level facial activities and their dynamics. This framework can also be seen as an information fusion process that combines the measurements from multiple levels. Finally, the expression, AU and facial features are recovered simultaneously through a probabilistic inference.

2.2 Eye Gaze Estimation

Currently, most eye gaze tracking algorithms (Morimoto and Mimica 2005; Guestrin and Eizenman 2006; Zhu and Ji 2007) and commercial gaze tracking products (<http://www.a-s-l.com>; Lc technologies 2005), are based on Pupil Center Corneal Reflection (PCCR) technique. One or multiple Infrared (IR) lights are used to illuminate the eye region and to build the corneal refraction (glint) on the corneal surface. At the same time, one or multiple cameras are used to capture the image of the eye. By detecting the pupil position and the glints in the image, the gaze direction can be estimated based on the relative position between the pupil and the glints. However, the gaze tracking systems based on IR

illumination have many limitations. First, the IR illumination can be affected by the sunshine in outdoor scenario. Second, the relative position between the IR lights and the camera need to be calibrated carefully. Third, because the pupil and the glint are very small, usually a high-resolution camera is needed. So, most current gaze tracking system can only work in indoor scenario.

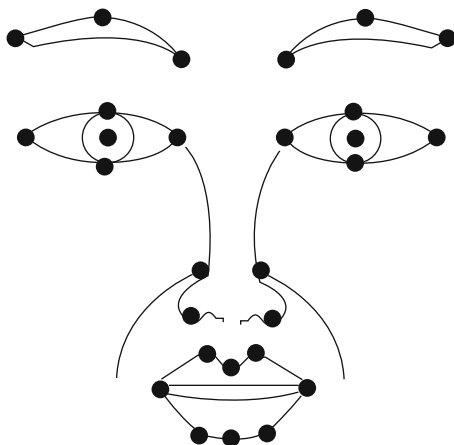
Here, based on the anatomical structure of the eyeball, we propose an extended 3D eye model, which includes the eyeball center, cornea center, pupil center, and facial feature points, for example, eye corners. By solving the equations of this 3D eye model, we can estimate the gaze from facial feature points. Our method can achieve accurate gaze estimation under free head movement.

3 Facial Feature Point Tracking

Facial feature point tracking is the essential component of our system, because it provides the basic measurements to the following face pose estimation, facial expression recognition, and eye gaze estimation. In this section, we track 28 facial feature points as shown in [Fig. 35.2](#). These points are located at well-defined positions such as eye corners, top points of eyebrows, and mouth corners.

3.1 Facial Feature Point Representation

For feature point detection, we capture the local information of each facial points using multi-scale and multi-orientation Gabor wavelets (Daugman 1988). Gabor-wavelet-based feature representation has the psychophysical basis of human vision and achieves robust



■ Fig. 35.2

Feature points are marked by *black circles* in the face model

performance for facial feature representation (McKenna et al. 1997; Jiao et al. 2003) under illumination and appearance variations.

For a given pixel $\mathbf{x} = (x, y)^T$ in a gray scale image I , a set of Gabor coefficients $J_j(\mathbf{x})$ is used to model the local appearance around the point. The coefficients $J_j(\mathbf{x})$ are resulted from convolutions of image $I(\mathbf{x})$ with the 2D Gabor wavelet kernels ψ_j , that is,

$$J_j(\mathbf{x}) = \sum \sum I(\mathbf{x}') \psi_j(\mathbf{x} - \mathbf{x}') \quad (35.1)$$

Here, kernel ψ_j is a plane wave restricted by a Gaussian envelope function:

$$\psi_j(\mathbf{x}) = \frac{k_j^2}{\sigma^2} \exp\left(-\frac{k_j^2 \mathbf{x}^2}{2\sigma^2}\right) \left[\exp(ik_j \cdot \mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right) \right] \quad (35.2)$$

with the wave vector

$$\mathbf{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_v \cos \varphi_u \\ k_v \sin \varphi_u \end{pmatrix} \quad (35.3)$$

where $k_v = 2^{-(v+1)}$ with $v = 0, 1, 2$ is the radial frequency in radians per unit length; $\varphi_u = \frac{\pi}{6}u$ with $u = 0, 1, \dots, 5$ is the wavelet orientation in radians, rotated counter-clockwise around the origin; $j = u + 6v$; and $i = \sqrt{-1}$ in this section. In this work, $\sigma = \pi$ is set for a frequency bandwidth of one octave.

Thus, the set of Gabor kernels consists of three spatial frequencies and six different orientations, and eighteen Gabor coefficients in the complex form are used to represent the pixel and its vicinity. Specifically, a jet vector \mathbf{J} is used to denote $(J_0, J_1, \dots, J_{17})$, where $J_j = a_j \exp(i\varphi_j)$, a_j and φ_j are the magnitude and phase of the j th Gabor coefficient. The Gabor wavelet jet vector is calculated for each marked fiducial point in training images. Given a new image, the fiducial points are searched by the sample jets from the training data. The similarity between two jet vectors is measured with the following phase-sensitive distance function:

$$D_\phi(\mathbf{J}, \mathbf{J}') = 1 - \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \mathbf{d} \cdot \mathbf{k}_j)}{\sqrt{\sum_j a_j^2 * \sum_j a'^2_j}} \quad (35.4)$$

where jet vectors \mathbf{J} and \mathbf{J}' refer to two locations with relative small displacement \mathbf{d} . The basic idea to estimate the displacement \mathbf{d} is from the Fourier shift property, that is, a shift \mathbf{d} in the spatial domain can be detected as a phase shift $\mathbf{k} \cdot \mathbf{d}$ in the frequency domain. In other words the phase change $\Delta\varphi$ is proportional to the displacement \mathbf{d} along the direction of the local frequency \mathbf{k} . The displacement between the two locations can be approximately estimated by minimizing the phase-sensitive distance $D_\phi(\mathbf{J}, \mathbf{J}')$ in **Eq. 35.4** as in (Fleet and Jepson 1990; Theimer 1994):

$$\mathbf{d}(\mathbf{J}, \mathbf{J}') = \begin{pmatrix} d_x \\ d_y \end{pmatrix} \approx \frac{1}{\Gamma_{xx} \Gamma_{yy} - \Gamma_{xy} \Gamma_{yx}} \times \begin{pmatrix} \Gamma_{yy} - \Gamma_{yx} \\ -\Gamma_{xy} \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} \quad (35.5)$$

$$\text{if } \Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx} \neq 0$$

where

$$\begin{aligned}\Phi_x &= \sum_j a_j a'_j k_{jx}(\phi_j - \phi'_j), \\ \Gamma_{xy} &= \sum_j a_j a'_j k_{jx} k_{jy},\end{aligned}$$

and Φ_y , Γ_{xx} , Γ_{yy} and Γ_{yx} are defined accordingly.

The phase-sensitive distance defined in [Eq. 35.4](#) changes rapidly with location, which helps accurately localize fiducial points in the image. Compensated by the displacement in [Eq. 35.5](#), the search of facial feature points can achieve subpixel sensitivity. This feature detection algorithm search for each feature points separately. In order to restrict the tracked feature points in a feasible face shape, we use the following ASM.

3.2 Pose-Dependent Active Shape Model

3.2.1 Active Shape Model

Given the facial feature points for a particular face view, a point distribution model can be constructed to characterize possible shape variations of human faces. Using the principle of the ASM, the point distribution model is constructed from a training set of face images. Facial feature points are marked on each face to outline its structure characteristics, and for each image a shape vector is used to represent the positions of feature points. All face shape vectors are aligned into a common coordinate frame by Procrustes transform (Dryden and Mardia 1998). Then the spatial constraints within feature points are captured by principal component analysis (PCA) (Cootes et al. 1995). A face shape vector s can be approximated by

$$s = \bar{s} + \mathbf{P}\mathbf{b} \quad (35.6)$$

where \bar{s} is the mean face shape; \mathbf{P} is a set of principal orthogonal modes of shape variation; and \mathbf{b} is a vector of shape parameters.

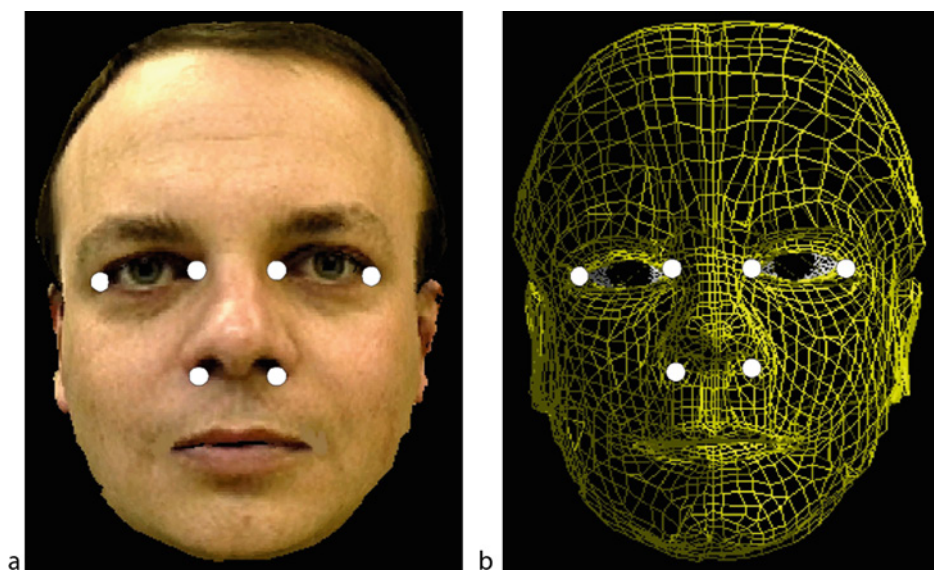
The face shape can be deformed by controlling the shape parameters. By applying limits to the elements of the shape parameter vector, it is possible to ensure that the generated shape is an admissible configuration of human face. The ASM approach searches the face shape by an iterative procedure. At each iteration the algorithm seeks to match each feature point locally based on its Gabor jet, and then refine all feature point locations by projecting them to the PCA shape space of the entire face. Fitting the entire face shape helps mitigate the errors in individual feature matchings.

3.2.2 Robust Face Pose Estimation

The face shape model, which we have introduced so far, basically assumes normative frontal face. The shape model will vary significantly, if face pose moves away from the frontal face. To compensate facial shape deformation due to face pose, we propose to estimate the 3D face pose and then use the estimated 3D pose to correct the shape model.

Given detected feature points at the previous frame, the 3D face pose can be efficiently estimated. In order to minimize the effect of facial expressions, only a set of rigid feature points that are not sensitive to facial expression changes is selected to estimate the face pose. Specifically, six feature points are selected, which include the four eye corners and two points at the nose's bottom shown in [Fig. 35.3a](#).

In order to estimate the face pose, the 3D shape model composed of these six facial features has to be initialized. Currently, the coordinates $\mathbf{X}_i = (x_i, y_i, z_i)^T$ of the six facial feature points in the 3D facial shape model are first initialized from a generic 3D face model as shown in [Fig. 35.3b](#). Due to the individual difference with the generic face model, the x and y coordinates of each facial feature point in the 3D face shape model are adjusted automatically to the specific individual based on the detected facial feature points in the initial frontal face view image.



■ Fig. 35.3

A synthesized frontal face image (a) and its 3D face geometry (b) with the rigid facial feature points marked by the *white dots*

Since the depth values of the facial feature points are not available for the specific individual, the depth pattern of the generic face model is used to approximate the z_i value for each facial feature point. Our experiment results show that this method is effective and feasible in our real-time application.

Based on the personalized 3D face shape model and these six detected facial feature points in a given face image, the face pose vector $\alpha = (\sigma_{\text{pan}}, \varphi_{\text{tilt}}, K_{\text{swing}})^T$ can be estimated accurately, where $(\sigma_{\text{pan}}, \varphi_{\text{tilt}}, K_{\text{swing}})$ are the three face pose angles and λ is the scale factor. Because the traditional least-square method (Or et al. 1998) cannot handle the outliers successfully, a robust algorithm based on RANSAC (Fischler and Bolles 1981) is employed to estimate the face pose accurately.

The pose estimation algorithm is briefly summarized as follows. The procedure starts with randomly selecting three feature points to form a triangle T_i . Under weak perspective projection model (Trucco and Verri 1998), each vertex (c_k, r_k) of T_i in the given image and the corresponding point (x_k, y_k) on the 3D face model are related as follows:

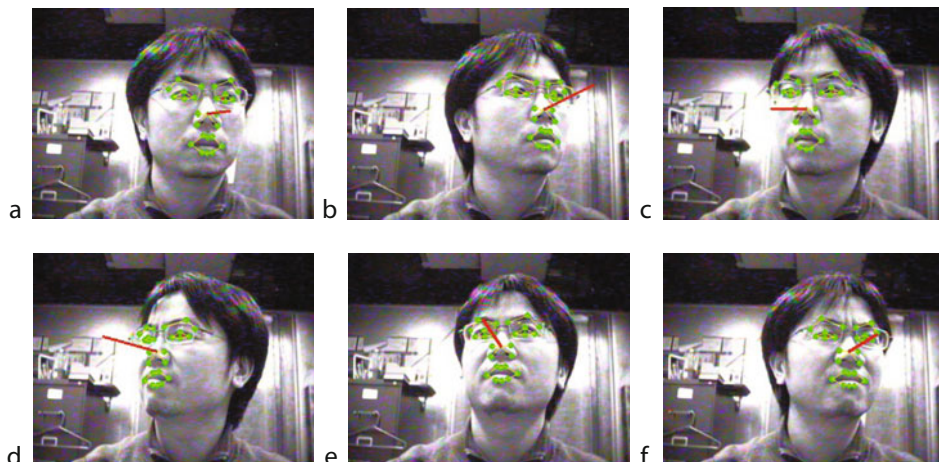
$$\begin{pmatrix} c_k - c_0 \\ r_k - r_0 \end{pmatrix} = \mathbf{M}_i \begin{pmatrix} x_k - x_0 \\ y_k - y_0 \end{pmatrix} \quad (35.7)$$

where $k = 1, 2$, and 3 ; \mathbf{M}_i is the projection matrix; (c_0, r_0) and (x_0, y_0) are the centers of the triangle in the given image and the reference 3D face model, respectively. Given the three detected feature points and the corresponding points on the 3D face model, we can solve the projection matrix \mathbf{M}_i for T_i . Using \mathbf{M}_i , the face model is projected onto the given image. A projection error e_i is then computed for all six feature points. e_i is then compared with a threshold e_0 , which is determined based on the amount of outliers estimated. \mathbf{M}_i is discarded, if e_i is larger than e_0 . Otherwise, a weight ω_i is computed as $(e_i - e_0)^2$ for \mathbf{M}_i . After repeating the above procedure for each triangle formed by the six feature points, we will get a list of matrices \mathbf{M}_i and their corresponding weights ω_i . From each projection matrix \mathbf{M}_i , a face pose vector α_i is computed uniquely after imposing some consistency constraints. Then the final face pose vector can be obtained as:

$$\alpha = \frac{\sum_{i=1}^K \alpha_i * \omega_i}{\sum_{i=1}^K \omega_i} \quad (35.8)$$

3.2.3 Face Shape Compensation

Given the estimated 3D face pose, the ASM is modified accordingly. Specifically, for each frame, the mean shape is modified by projecting it to the image plane using the estimated face pose through Eq. 35.7. The modified mean shape is more suitable for the current pose and provides better shape constraint for the feature point search. Moreover, the projected mean shapes offer good initialization to avoid being trapped into local minima during the feature search process.



■ Fig. 35.4

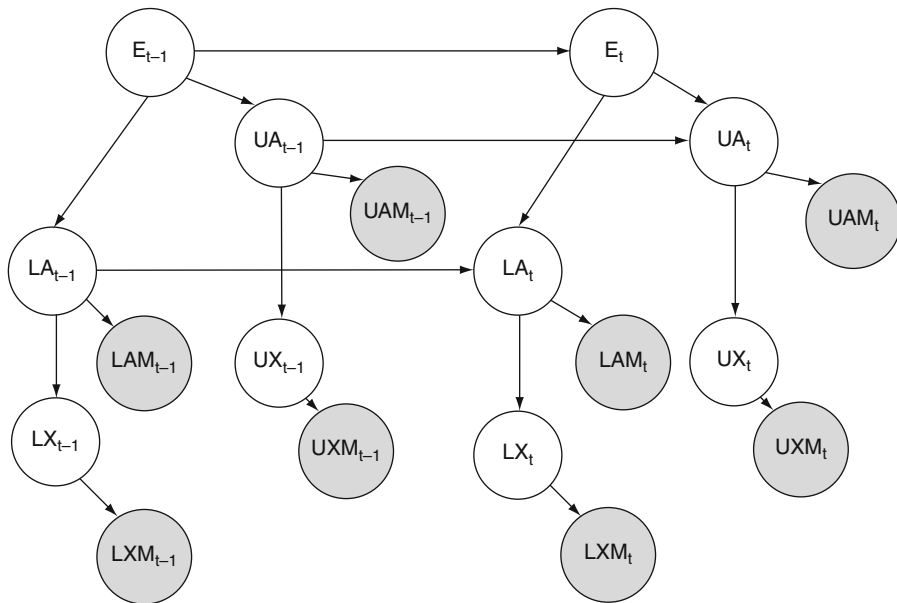
The face pose estimation and facial feature tracking results: face normal is represented by the line, and the detected facial feature points are marked by dots

The face pose estimation and facial feature tracking results are shown in ► Fig. 35.4. The face normal is perpendicular to the face plane and represented by the three estimated Euler face pose angles.

4 A Unified Framework for Simultaneous Facial Activity Tracking

The facial feature point tracking algorithm, which we have introduced so far, does not consider facial deformation due to large expression change. For example, it is difficult for ASM to capture the mouth shape of yawning, because it is quite different from the mouth shape in neutral facial expression. In this section, we propose to use a unified DBN to capture the relationships between facial feature points and facial expressions, and to track and recognize them simultaneously.

We propose to use the DBN model in ► Fig. 35.5 to track the three levels of facial activities. The E_t node in the top level represents the current expression; UA_t and LA_t represent the AUs related to the upper-face and the AUs related to the lower-face, respectively; UX_t and LX_t node denotes the facial feature points on the upper-face and the facial feature points on the lower-face, respectively. UAM_t , LAM_t , UXM_t and LXM_t represent the corresponding measurements of AUs and facial feature points. Here, the measurement of facial feature points is extracted through the introduced ASM-based facial feature tracker. The measurement of each AU is obtained by the Adaboost classifier similar to Bartlett et al. (2006).



■ Fig. 35.5

The DBN model for simultaneous facial feature tracking and expression recognition

Given the measurement sequences, the posterior of the three level facial activities are estimated through the inference in DBN (► Sect. 4.2). And the optimal states are tracked by maximizing this posterior:

$$E_t^*, UA_t^*, LA_t^*, UX_t^*, LX_t^* = \arg \max_{E_t, UA_t, LA_t, UX_t, LX_t} P(E_t, UA_t, LA_t, UX_t, LX_t | UAM_{1:t}, LAM_{1:t}, UXM_{1:t}, LXM_{1:t}) \quad (35.9)$$

4.1 DBN Model Parameterization

Given the DBN model in ► Fig. 35.5, we need to define the states for each node and the conditional probability distribution (CPD) associated with each node. The CPD defines conditional probability of each node given its parents $P(X|pa(X))$. Hereafter, $pa(X)$ is defined as the set of parent nodes of node X . In this section, we will define the CPD of each node.

4.1.1 Expression Level

In the top expression level, we want to model the six basic expressions. E_t is a discrete node which has eight possible states: happiness, sadness, disgust, surprise, anger, fear, neutral,

and “others.” The “others” state denotes all the expressions that cannot be explained by the basic expressions.

The CPD of the expression $P(E_t|E_{t-1})$ can be represented as a 8×8 transition matrix T whose entries $T_{e,e'}$ denotes the probability of the transition from expression e to e' . Although this matrix can be learned from training data, we have found that a general near-diagonal matrix works well for all the sequences. We set the diagonal elements to be close to one, and the rest of the percentages are equally distributed for other expressions. This matrix actually gives a higher probability to the current expression if it is same as the previous one.

4.1.2 AU Level

The six typical expressions only describe the global facial activities, and they are only a small set of the complex face expressions. For example, the “surprise” implies a widely opened mouth and a raise of the eyebrows. But in applications, the subject may open mouth widely without raising eyebrows.

It is generally believed that the expressions can be described linguistically using culture and ethnically independent AUs. Such AUs were developed by Ekman and Friesen in their FACS (Ekman and Friesen 1978), where each AU is coded based on the local facial muscle involvements. For example, AU27 (mouth stretch) describes a widely open mouth, and AU4 (brow lower) makes the eyebrows lower and pushed together. In Zhang and Ji (2005), they exploit some primary AUs which are highly related to the basic facial expressions. Here, we model 11 AUs from these primary AUs. They are listed in Fig. 35.6.

Based on our DBN model in Fig. 35.5, we first group the AUs into “upper-face AUs (UA)”, including AU1, AU2, AU4, AU6, and “lower-face AUs (LA)”, including AU12, AU15, AU17, AU23, AU24, AU25, and AU27. The two AU groups capture the local facial behaviors of upper-face and low-face, respectively.












AU1  Inner Brow Raiser	AU2  Outer Brow Raiser	AU4  Brower Lowerer	AU6  Cheek Raiser	AU12  Lip Corner Puller	AU15  Lip Corner Depressor
AU17  Chin Raiser	AU23  Lip Tightener	AU24  Lip Pressor	AU25  Lips part	AU27  Mouth Stretch	

Fig. 35.6
The list of AUs

Each single AU has two discrete values: 0 and 1, which represents “presence” and “absence,” respectively. However, if we directly stack these binary AUs into a vector and model the state of the node UA (or LA) with this vector, there will be too many possible states. For example, for the lower face AU, there will be 2^7 possible states, but most of them rarely occur in daily life, that is, have too few examples in the training data to learn their probabilities. In this work, we select a few most frequent AU combinations as typical AUs, and use the typical AUs as the states of AU node. Specifically, LA has eight states including AU25, AU25 + AU12, AU25 + AU27, AU12, AU17/AU15, AU23/AU24, “neutral,” and “others” and UA has six states including AU6, AU6 + AU4, AU1 + AU2, AU1 + AU4, “neutral,” and “others.” (“+” means two AUs happen together, “/” means either of the two AUs happens.) Here, the “others” state denotes all the AU combinations that cannot be explained as typical AUs.

Finally, the measurement nodes (LAM and UAM) for the AUs represent their observations obtained through the AU classifier in Bartlett et al. (2006). The AU measurement has the same discrete states as its corresponding AU. So, the conditional probability $P(LAM|LA)$ is modeled as a conditional probability table (CPT), which is an 8×8 matrix. Similarly, the CPT of $P(UAM|UA)$ is a 6×6 matrix. And these conditional probabilities represent the measurement uncertainty with AU classifier.

4.1.3 Facial Feature Level

In this work, we focus on the facial feature points around the mouth and eyes, which have significant movement under different expressions. Because the eye and mouth shapes in neutral face are different for different subject, to eliminate the neutral shape variance we subtract the neutral shape from current shape, and model the shape difference. LX is a 16 dimensional vector which denotes the x,y differences of the eight mouth points. Similarly, UX denotes differences of the 16 eye and eyebrow points.

Given the local AU, the CPD of facial feature points can be represented as a Gaussian distribution, for example, for lower-face:

$$P(LX_t|LA_t = k) = N(LX_t; \mu_k, \Sigma_k) \quad (35.10)$$


with the mean shape vector μ_k and covariance matrix Σ_k . Based on the conditional independence embedded in the BN, we could learn μ_k and Σ_k locally from training data.

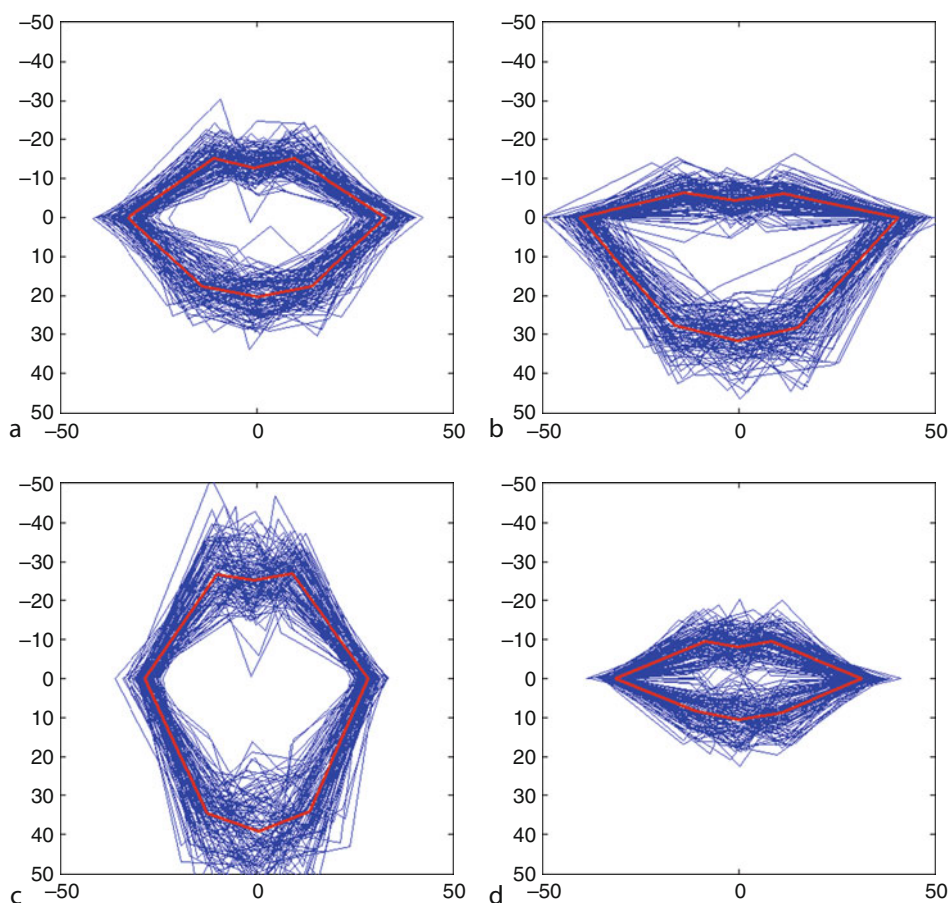
Finally, the measurement nodes of the feature points represent their positions extracted from ASM-based tracker. The facial feature measurements are the continuous vectors, which have the same dimension as their parents. And CPD of measurement is modeled as a linear Gaussian distribution (Murphy 1998), for example, for lower face:


$$P(LXM_t|LX_t = lx) = N(LXM_t; W_L \cdot lx + \mu_L, \Sigma_L) \quad (35.11)$$

with the mean shape vector μ_L , regression matrix W_L , and covariance matrix Σ_L . These parameters can be learned from the training data.

Given the definition of the CPDs in the DBN model, we learn the parameters of the CPDs from Cohn and Kanade's DFAT-504 facial expression database (Kanade et al. 2000), which includes 486 sequences from 97 subjects.

For example, in the facial feature level,  Fig. 35.7 shows 200 samples drawn from the learned CPDs of the lower-face facial feature points: $P(LX_t|LA_t)$. (The LX_t in our model is shape difference. For clarity, we show the distribution of LX_t by adding a constant neutral shape: $P(LX_t + C|LA_t)$, where C is a constant neutral shape.) We can see that the local facial feature distributions for different AUs are different. Thus, the AU actually can provide a prior probability of the local shape.



 Fig. 35.7

The CPDs of low-face facial feature points

4.2 DBN Inference: Simultaneous Facial Feature Tracking and Expression Recognition

Given the DBN model and the measurements of facial features, AUs, and expressions, we want to simultaneously localize the feature points, and recognize AUs and the facial expressions by maximizing the posterior probability of the hidden nodes as [Eq. 35.9](#). Based on the DBN model, the posterior probability of the expression, AUs and facial features can be computed from the posterior of the previous frame:

$$\begin{aligned}
 & P(E_t, UA_t, LA_t, UX_t, LX_t | UAM_{1:t}, LAM_{1:t}, UXM_{1:t}, LXM_{1:t}) \\
 & \propto P(UAM_t | UA_t) P(LAM_t | LA_t) P(UXM_t | UX_t) P(LXM_t | LX_t) \\
 & P(UX_t | UA_t) P(LX_t | LA_t) \int_{E_{t-1}, UA_{t-1}, LA_{t-1}} P(E_t | E_{t-1}) \\
 & P(UA_t | UA_{t-1}, E_t) P(LA_t | LA_{t-1}, E_t) \\
 & P(E_{t-1}, UA_{t-1}, LA_{t-1} | UAM_{1:t-1}, LAM_{1:t-1}, UXM_{1:t-1}, LXM_{1:t-1})
 \end{aligned} \quad (35.12)$$

This filtering problem in DBN can be solved by “Interface algorithm” (Murphy 2002) efficiently.

4.3 Result on the Cohn-Kanade Database

We test our method on Cohn and Kanade’s DFAT-504 database (C-K database) (Lien et al. 1999), which includes 97 subjects covering different races, ages, and genders. For comparison, we implemented Bartlett et al.’s method (Bartlett et al. 2005) as our baseline AU recognition system. In the experiment reported on their website (Bartlett et al. 2007), they test on 313 frames of peak AU (i.e., highest magnitude of the target expression) and 313 frames of neutral expression. Using leave-one-subject-out cross validation, the overall recognition rate is 93.6%. In our experiments, we test on 5,070 frames from 463 sequences in C-K database, including peak AUs, neutral expressions, and weak AUs (low magnitude of target expression). We also use the leave-one-subject-out cross validation to evaluate our baseline system on 11 AUs, as shown in [Fig. 35.6](#). It achieves 91.77% recognition rate, with 80.52% TPR and 5.35% FPR.

The above experiments are focused on classify the binary state (presence/absence) of each AU. However, some AUs can be combined to represent different expressions, for example, AU25 + AU27 and AU25 + AU12 in [Fig. 35.7](#) represent different mouth expressions. Our method is focused on recognizing these AU combinations. As discussed before, our model classifies each frame into one of six upper-face AU combinations and one of eight lower-face AU combinations. For this challenging multi-class classification problem, the baseline system achieves classification rate of 69.74% for upper-face AU and 66.41% for lower-face AU.

■ Table 35.1

Comparison with start-of-the-art AU recognition and facial feature tracking systems on C-K Database. (Notice that only the proposed method can track different facial activities, i.e., facial feature points, AU and expression, simultaneously)

	Upper-face AU (UA)	Lower-face AU (LA)	Facial feature points	Expression
Proposed method	70.34%	68.88%	1.89 pixels	60.85%
AU baseline (Bartlett et al. 2005)	69.74%	66.41%	–	–
Feature point tracking baseline (Tong et al. 2007c)	–	–	1.99 pixels	–

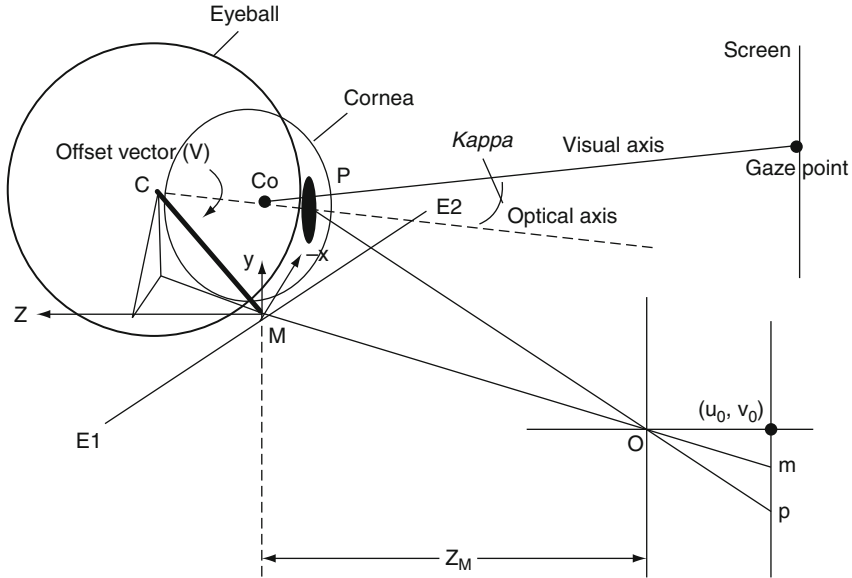
To evaluate our facial feature point tracking result, we also compare with the ASM-based facial feature tracker in ▶ Sect. 4, which achieves average tracking error (mean square error) of 1.99 pixels.

We summarize the results of our proposed methods and the baseline systems in ▶ Table 35.1. We can see that our model can improve both the AU recognition and facial feature tracking results. Besides more accurate facial feature tracking and AU recognition, our model can recognize eight global expressions with recognition rate of 60.85%.

5 3D Gaze Estimation Based on Facial Feature Points Without IR Illumination

In this section, we introduce the gaze estimation method based on the tracked facial feature points. The key of our algorithm is to compute the 3D position of the eyeball center C based on the middle point M of two eye corners (E_1, E_2) (▶ Fig. 35.8). The 3D model is based on the anatomical structure of the eye (Oyster 1999; Guestrin and Eizenman 2006). As shown in ▶ Fig. 35.8, the eyeball is made up of the segments of two spheres with different sizes. The anterior smaller segment is the cornea. The cornea is transparent, and the pupil is inside the cornea. The optical axis is defined as the 3D line connecting the corneal center C_0 and the pupil center P . Since the person's gaze (line-of-sight) is defined as the visual axis rather than the optical axis, the relationship between these two axes has to be modeled. The angle between the optical axis and the visual axis is named as $kappa$, which is a constant value for each person.

When the person gazes at different directions, the corneal center C_0 and the pupil center P will rotate around the eyeball center C , and C_0, P, C are all on the optical axis. Since C is inside the face, we have to estimate its position from the facial point on the face surface: E_1 and E_2 are the two eye corners and M is their middle point. The offset vector V



■ Fig. 35.8
3D eye model

between **M** and **C** is related to the face pose. Based on the eye model, we can estimate the gaze step by step as follows.

5.1 Step 1. Facial Feature Point Tracking and Face Pose Estimation

First, based on the facial feature point tracking algorithm in 🔗 Sect. 4 and the face pose estimation algorithm in 🔗 Sect. 3.2.2, we can track the facial points and estimate the face pose vector $\alpha = (\sigma_{\text{pan}}, \varphi_{\text{tilt}}, \kappa_{\text{swing}}, s)$, where $(\sigma_{\text{pan}}, \varphi_{\text{tilt}}, \kappa_{\text{swing}})$ are the three face pose angles and s is the scale factor.

Here, the face pose is estimated through a generic 3D face model, which is composed of six rigid face points (🔗 Fig. 35.9). Actually, the three face pose angles can define a 3×3 rotation matrix R to rotate the 3D face point from the face-model coordinate to the camera coordinate. Assuming weak perspective projection, the projection from 3D point in face-model coordinate to the 2D image point is defined as

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = sR_{1,2} \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} + \begin{pmatrix} u_0^f \\ v_0^f \end{pmatrix} \quad (35.13)$$

$R_{1,2}$ is a 2×3 matrix which is composed of the first two rows of the rotation matrix R . (u_0^f, v_0^f) is the projection of the face-model origin. (Here, the face origin is defined as the nose tip and the z axis is pointing out the face.)

5.2 Step 2. Estimate the 3D Model Point M

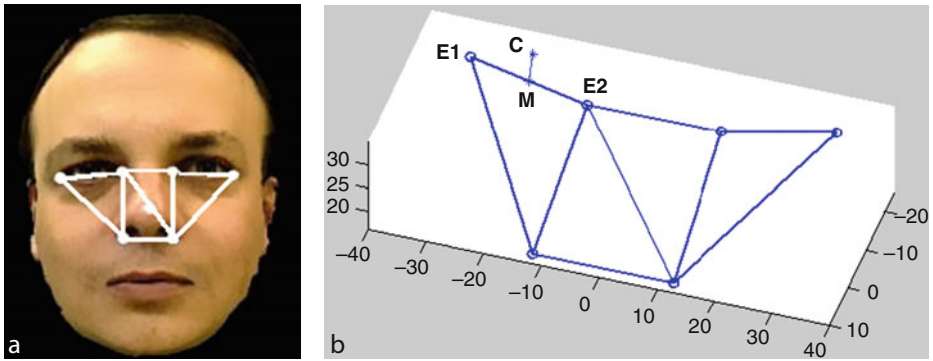
To estimate the 3D eyeball center C , we extend the face model by adding two points (C and M) in the 3D face model, as shown in [Fig. 35.9](#). So, although the camera cannot capture the eyeball center C directly, we can first estimate the 3D point M from the tracked facial feature points, and then estimate the C position from M . Our implicit assumption here is that the relative spatial relationships between M and C is fixed independent of gaze direction and head movement.

As shown in [Fig. 35.8](#), from the tracked 2D eye corner points $E1$ and $E2$, the eye corner middle point in the image is estimated $\mathbf{m} = (u_m, v_m)^T$. If we can estimate the distance z_M from 3D point M to the camera, the 3D coordinates of M can be recovered.

Same as the work by Ishikawa et al. (2004) 3D distance from the camera can be approximated using the fact that the distance is inversely proportional to the scale factor (s) of the face: $z_M = \frac{\lambda}{s}$. Here, the inverse-proportional factor λ can be recovered automatically in the user-dependent calibration procedure.

After the distance z_M is estimated, M can be recovered using [Eq. 35.13](#):

$$\mathbf{M} = \frac{z_M}{f} \begin{pmatrix} u_m - u_0 \\ v_m - v_0 \\ f \end{pmatrix} \quad (35.14)$$



■ Fig. 35.9

3D generic face model. (a) A frontal face image and the selected rigid points and triangles. (b) 3D face model with middle point and eyeball center

where f is the camera focus length, $(u_0, v_0)^T$ is the projection of the camera origin, as shown in ② Fig. 35.8. Assuming we use a calibrated camera, f and $(u_0, v_0)^T$ are, therefore, known.

5.3 Step 3. Compute 3D Position of C

We compute the eyeball center **C** based on the middle point **M** and the offset vector **V** in ② Fig. 35.8. Here, the offset vector **V** in the camera frame is related to the face pose.

Since the middle point **M** and the eyeball center point **C** are fixed relative to the 3D face model (② Fig. 35.9). Their position in this face-model coordinate are $M^f = (x_M^f, y_M^f, z_M^f)^T$ and $C^f = (x_C^f, y_C^f, z_C^f)^T$. Let the rotation matrix and the translation between the face coordinate and the camera coordinate be **R** and **T**, so the offset vector between **M** and **C** in camera coordinate is:

$$\begin{aligned} & \mathbf{C} - \mathbf{M} \\ &= (R\mathbf{C}^f + \mathbf{T}) - (R\mathbf{M}^f + \mathbf{T}) \\ &= R(\mathbf{C}^f - \mathbf{M}^f) \\ &= R\mathbf{V}^f \end{aligned} \quad (35.15)$$

where $\mathbf{V}^f = \mathbf{C}^f - \mathbf{M}^f$ is a constant offset vector in the face model, independent of gaze direction and head position. During tracking, given the face pose **R**, and the **M** position in camera coordinate, **C** in camera coordinate can be written as:

$$\mathbf{C} = \mathbf{M} + R\mathbf{V}^f \quad (35.16)$$

5.4 Step 4. Compute P

Given **C** and the pupil image $\mathbf{p} = (u_p, v_p)^T$, the 3D pupil position $\mathbf{P} = (x_P, y_P, z_P)^T$ can be estimated from its image coordinates and using the assumption that the distance between **C** and **P** is a constant *K*. Specifically, **P** can be solved using the following equations.

$$\begin{cases} \begin{pmatrix} u_p \\ v_p \end{pmatrix} = \frac{f}{z_P} \begin{pmatrix} x_P \\ y_P \end{pmatrix} + \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \\ \|\mathbf{P} - \mathbf{C}\| = K \end{cases} \quad (35.17)$$

5.5 Step 5. Compute Gaze

Since the distance (K_0) between the corneal center and the pupil center is also a constant, given the **P** and **C**, the corneal center **C**₀ can be estimated as:

$$C_0 = C + \frac{K_0}{K}(P - C) \quad (35.18)$$

Then, same as the work by Guestrin and Eizenman (2006), the visual axis can be obtained by adding two person-specific angles to the optical axis as follows:

$$V_{\text{vis}} = f(\alpha, \beta; P - C_0) \quad (35.19)$$

Here, $f(\alpha, \beta; \mathbf{V})$ is a function to add the horizontal angle (α) and vertical angle (β) to a vector \mathbf{V} .

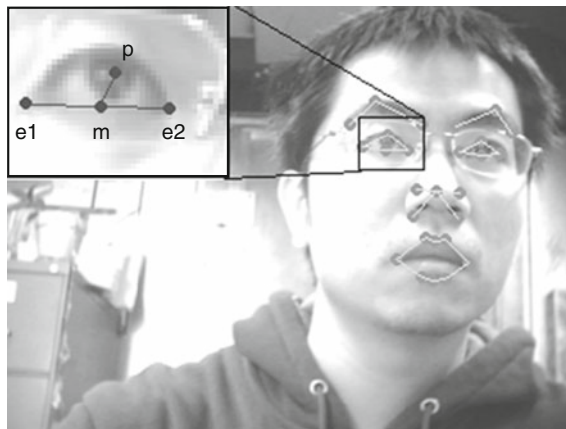
5.6 Gaze Estimation Result

In this preliminary experiment, we estimate the subject's gaze point on the monitor. The facial feature tracking result to estimate face pose and the eye features to estimate gaze are shown in [Fig. 35.10](#).

Then we first test the gaze estimation accuracy without head movement. The subject keeps his head still and gazes at nine points on the screen sequentially.

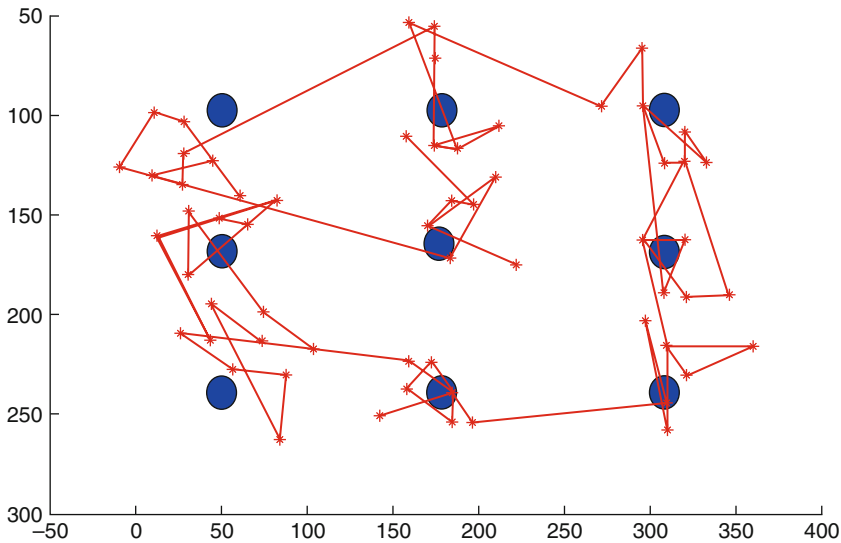
The estimated gazes (scan pattern) are shown in [Fig. 35.11](#). The accuracy can achieve: X accuracy = 17.7 mm (1.8288°), Y accuracy = 19.3 mm (2.0°).

Finally, we test our algorithm under free head movement. The head moves in the region of approximately $140 \times 140 \times 220$ mm (width \times height \times depth). The estimated gaze accuracy can achieve X accuracy = 22.42 mm (2.18°), Y accuracy = 26.17 mm (2.53°). We can see that our algorithm can give reasonable gaze estimation result ($<3^\circ$), and the accuracy does not decrease much with free head movement.



■ Fig. 35.10

Facial Feature tracking result and the points to estimate gaze



■ Fig. 35.11

Gaze estimation result. The large dark (gray) circles are the nine points showed on the screen. The stars and the lines denote the estimated gaze points and the saccade movement

6 Conclusion

This chapter introduces a computer vision system to track various facial behaviors which can be used to monitor the driver fatigue. Using one camera, this system can simultaneously track facial feature points, recognize facial expression, and estimate the eye gaze. The main contributions of our methods are summarized as follows:

1. To robustly track facial features under different face poses, we propose a pose-dependent ASM. We first estimate the 3D face pose through a robust pose estimation algorithm. Then the mean shape of ASM is modified based on the estimated face pose in every frame. This modified ASM provides better shape constraint for the feature point search.
2. We propose a unified framework to capture the relationships among facial feature points and facial expression. Then, through a simultaneous facial feature point tracking and facial expression recognition, both the tracking and recognition results can be improved.
3. Based on the tracked facial feature points, we introduce a Gaze Estimation Algorithm without IR Illumination, which is more suitable for Outdoor Gaze Estimation.

Our system monitors a wide range of facial behaviors, from detailed facial feature point movement, local facial movement, face pose, gaze movement, to global facial expression. From these visual cues, we can extract various parameters to characterize driver state. Study in Ji et al. (2006) discusses various such parameters as well as provides

a framework to integrate these parameters to achieve a robust characterization and prediction of driver state. We will pursue this study to develop a prototype driver state monitoring system.

References

- Anon (1998) Proximity array sensing system: head position monitor/metric. Advanced Safety Concepts, Sante Fe
- Anon (1999) Perclos and eyetracking: challenge and opportunity. Technical Report, Applied Science Laboratories, Bedford
- Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: Proceedings of IEEE Computer society conference computer vision and pattern recognition, San Diego
- Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2006) Automatic recognition of facial actions in spontaneous expressions. *J Multimedia* 1:22–35
- Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2007) <http://mplab.ucsd.edu/grants/project1/research/fully-auto-facs-coding.html>
- Bazzo J, Lamar M (2004) Recognizing facial actions using gabor wavelets with neutral face average difference. In: Proceedings of the sixth IEEE international conference on automatic face and gesture recognition, Seoul
- Bourel F, Chibelushi CC, Low AA (2000) Robust facial feature tracking. In: Proceedings of the British Machine Vision Conference, Bristol
- Boverie S, Lcquellec J, Hirl A (1998) Intelligent systems for video monitoring of vehicle cockpit. International congress and exposition ITS: advanced controls and vehicle navigation systems, pp 1–5
- Buenaposada JM, Munoz E, Baumela L (2008) Recognising facial expressions in video sequences. *Pattern Anal Appl* 11(1):101–116
- Cohen I, Sebe N, Ashutosh G, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. *Comput Vis Image Underst* 91(1–2):160–187
- Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models – their training and application. *Comput Vis Image Underst* 61(1):38–59
- Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance model. *IEEE Trans Pattern Anal Mach Intell* 23(6):681–685
- Daugman J (1988) Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Trans ASSP* 36(7):1169–1179
- Dinges D, Mallis M, Maislin G, Powell J (1998) Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Department of Transportation Highway Safety Publication 808 762
- Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Classifying facial actions. *IEEE Trans Pattern Anal Mach Intell* 21:974–989
- Dornaika F, Davoine F (2008) Simultaneous facial action tracking and expression recognition in the presence of head motion. *Int J Comput Vis (IJCV)* 76:251–281
- Dryden IL, Mardia KV (1998) Shape analysis. Wiley, Chichester
- Ekman P, Friesen WV (1978) Facial action coding system (FACS): manual. Consulting Psychologists Press, Palo Alto
- Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 24(6):381–395
- Fleet DJ, Jepson AD (1990) Computation of component image velocity from local phase information. *Int J Comput Vis* 5(1):77–104
- Guestrin ED, Eizenman M (2006) General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans Biomed Eng* 53:1124–1133
- Ishii T, Hirose M, Iwata H (1987) Automatic recognition of drivers facial expression by image analysis. *J Soc Automot Eng Japan* 41:1398–1403
- Ishikawa T, Baker S, Matthews I, Kanade T (2004) Passive driver gaze tracking with active

- appearance models. In: Proceedings of the 11th World congress on intelligent transportation systems, Berkeley
- Ji Q, Zhu Z, Lan P (2004) Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Trans Veh Technol* 53(4):1052–1068
- Ji Q, Lan P, Looney C (2006) A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Trans Syst Man Cybernetics Part A Syst Humans* 36(5):862–875
- Jiao F, Li SZ, Shum HY, Schuurmans D (2003) Face alignment using statistical models and wavelet features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR03), vol 1. Madison, pp 321–327
- Kanade T, Cohn J, Tian YL (2000) Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE international conference on automatic face and gesture recognition (FG'00), Grenoble
- Lc Technologies (2005) <http://www.eyegaze.com>
- Lien JJJ, Kanade T, Cohn J, Li C (1999) Detection, tracking, and classification of action units in facial expression. *J Robot Auton Syst* 31: 131–146
- McKenna SJ, Gong S, Würtz RP, Tanner J, Banin D (1997) Tracking facial feature points with gabor wavelets and shape models. In: Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication, Hilton Rye Town, pp 35–42
- Morimoto CH, Mimica MR (2005) Eye gaze tracking techniques for interactive applications. *Comput Vis Image Underst* 98:4–24 (Special issue on eye detection and tracking)
- Murphy K (1998) Inference and learning in hybrid bayesian networks. Report No. UCBCSD-98-990, Computer Science Department, U.C. Berkeley
- Murphy K (2002) Dynamic Bayesian networks (draft)
- Or SH, Luk WS, Wong KH, King I (1998) An efficient iterative pose estimation algorithm. *Image Vision Comput* 16(5):353–362
- Oyster CW (1999) The human eye: structure and function. Sinauer Associate, Sunderland
- Saito S (1992) Does fatigue exist in a quantitative of eye movement? *Ergonomics* 35:607–615
- Saito H, Ishiwaka T, Sakata M, Okabayashi S (1994) Applications of drivers's line of sight to automobiles – what can driver's eye tell. In: Proceedings of 1994 vehicle navigation and information systems conference, Yokohama, pp 21–26
- Shan C, Gong S, MacOwan PW (2006) Dynamic facial expression recognition using a bayesian temporal manifold model. In: Proceedings of the British machine vision conference, Edinburgh
- Smith P, Shah M, da Vitoria Lobo N (2000) Monitoring head/eye motion for driver alertness with one camera. In: Proceedings of the 15th international conference on pattern recognition, vol 4. Barcelona, pp 636–642
- Theimer WM (1994) Phase-based binocular vergence control and depth reconstruction using active vision. *CVGIP: Image Underst* 60(3):343–358
- Tian YI, Kanade T, Cohn J (2001) Recognizing action units for facial expression analysis. *IEEE Trans Pattern Anal Mach Intell* 23:33–80
- Tomasi C, Kanade T (1991) Detection and tracking of point features. Carnegie Mellon University Technical Report CMU-CS-91-132
- Tong Y, Liao W, Ji Q (2007a) Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans Pattern Anal Mach Intell* 29:1683–1699
- Tong Y, Liao W, Xue Z, Ji Q (2007) A unified probabilistic framework for spontaneous facial activity modeling and understanding. In: Proceedings of the 2007 IEEE Conference on computer vision and pattern recognition (CVPR), Minneapolis
- Tong Y, Wang Y, Zhu Z, Ji Q (2007c) Robust facial feature tracking under varying face pose and facial expression. *Pattern Recogn* 40: 3195–3208
- Tong Y, Chen J, Ji Q (2010) A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Trans Pattern Anal Mach Intell* 32(2):258–273
- Trucco E, Verri A (1998) Introductory techniques for 3-D computer vision. Prentice-Hall, Upper Saddle River
- Wiskott L, Fellous JM, Krger N, von der Malsburg C (1997) Active appearance model. *IEEE Trans Pattern Anal Mach Intell* 19:775–779
- Yamamoto K, Higuchi S (1992) Development of a drowsiness warning system. *J Soc Automot Engin Japan* 46:127–133
- Zhang Y, Ji Q (2005) Active and dynamic information fusion for facial expression understanding from

- image sequences. IEEE Trans Pattern Anal Mach Intell (PAMI) 27(5):699–714
- Zhou S, Krueger V, Chellappa R (2003) Probabilistic recognition of human faces from video. Comput Vis Image Underst 91(1–2):214–245
- Zhu Z, Ji Q (2007) Novel eye gaze tracking techniques under natural head movement. IEEE Trans Biomed Eng 54(12):2246–2260
- <http://www.a-s-l.com> (2006)

36 Drowsy and Fatigued Driving Problem Significance and Detection Based on Driver Control Functions

Azim Eskandarian¹ · Ali Mortazavi² · Riaz Akbar Sayed³

¹Center for Intelligent Systems Research, The George Washington University, Washington, DC, USA

²Partners for Advanced Transportation Technology (PATH), University of California, Berkeley, CA, USA

³Mechanical Department, NWFP University of Engineering and Technology, Peshawar, North West Frontier Provi, Pakistan

1	<i>Introduction and Problem Statement</i>	943
2	<i>Characteristics of Drowsiness and Fatigue</i>	944
3	<i>Characteristics of Drowsiness and Fatigue-Related Accidents</i>	946
4	<i>Drowsiness and Fatigue Detection Systems</i>	948
5	<i>Sensing Driver's Physical and Physiological Conditions</i>	948
6	<i>Sensing Vehicle State Variables</i>	949
7	<i>Correlation between Drowsiness and Vehicle State Variables</i>	949
7.1	Vehicle Steering Activity	950
7.1.1	Frequency of Steering Wheel	951
7.1.2	Steering Wheel Reversal Rate	951
7.1.3	Steering Correction	951
7.1.4	Steering Velocity	953
7.1.5	Steering Amplitude Duration Squared Theta	953
7.1.6	Weight Flat Zero	954

7.2 Lateral Position 954

7.3 Vehicle Speed 954

7.4 Yaw/Brake/Acceleration 955

8 *Drowsiness Detection Methods Based on Vehicle State Variables*955

9 *Examples of Drowsiness Detection Methods* 956

10 *Advantages and Disadvantages of Detection Systems Using Vehicle State Variables* 967

11 *Design and Implementation Challenges to Consider* 968

12 *Conclusion* 970

Abstract: Drowsy and fatigue driving is a major transportation safety concern and is responsible for thousands of accidents and numerous fatalities every year. The resulting harms of drowsy/fatigue driving could be even higher among commercial vehicles. Drowsy driving crashes are usually of high severity due to the drivers' significant loss of control, often leading to unpredicted vehicle trajectory and no braking response. Reliable safety systems are needed to mitigate these crashes. The most important challenge is to detect the driver's condition sufficiently early, prior to the onset of sleep, to avoid collisions.

Various detection methods have been proposed by researchers and a few systems are available in the commercial market. In general, drowsiness detection methods fall into two major categories of monitoring physiological and physical conditions of the drivers and monitoring vehicle-related variables based on driver control functions that correlate with the driver's level of drowsiness. Each method has its advantages and shortcomings. A reliable detection method needs to be integrated with a safety system which may include advisory warning, semi-control, or full control of vehicle, i.e., braking and steering to achieve safe conditions. The type and intensity of warning or control should also be carefully selected and are discussed in another chapter.

This chapter first reviews the statistical significance of the crash data due to drowsiness and fatigue conditions. Then, the issues concerning various detection methods are discussed. Detection systems based on driver control functions are mainly discussed in this chapter. The concepts and approaches presented in this section are from a comprehensive literature review including the author's past research; they can guide the development of safety systems for a passenger or commercial vehicles.

1 Introduction and Problem Statement

In the last 10 years (2000–2010), more than 12,500 people have lost their lives in crashes related to driver fatigue/drowsiness (Fatality Analysis Reporting System (FARS) of the US department of transportation). Driver fatigue is particularly significant in commercial vehicle operations because truck drivers stay on the road for extended period of time, which often involves driving at night. Surveys conducted by Tilley in 1973 at Duke University, Seko in 1984 in Japan, and Planque in 1991 in France all indicate driver drowsiness as a serious problem and major cause of fatal accidents. According to a survey by the American Automobile Association (AAA) Foundation in 2004, 90% of the North American police officers have stopped a driver who they believed was drunk, but turned out to be drowsy (aaaafoundation.org). According to some estimates drowsy driving costs the North American consumer \$16.4 billion in terms of property damages, health claims, lost time and productivity. Another \$60.4 billion/year are spent by US Government and businesses on accidents related to drowsy driving.

The Fatality Analysis Reporting System (FARS 2008) contains annually updated census data for all fatal crashes occurring within the United States. The National Automotive Sampling System–General Estimates System (NASS/GES 2008) is a nationally representative sample of police-reported crashes occurring within the United States.

Year	Number of Fatal Crashes	Number of Fatalities
1999	1352	1564
2000	1342	1538
2001	1214	1362
2002	1237	1418
2003	1083	1233
2004	1131	1298
2005	1017	1175
2006	978	1072
2007	908	1031
2008	719	827
Total	10981	12518

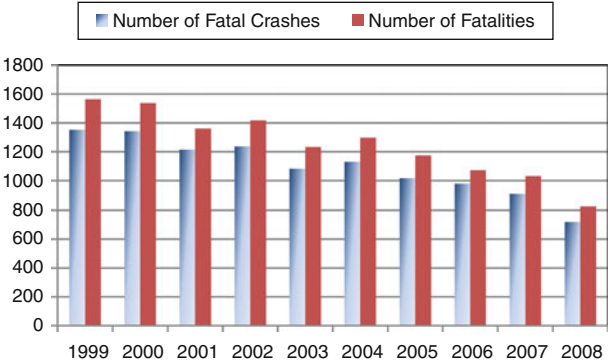


Fig. 36.1
Number of fatal crashes and fatalities in crashes involving a vehicle in which a “Sleepy, fatigued, or drowsy driver” was documented as a contributing factor over a 10-year period

Each crash in the sample is assigned a weighting factor that can be used to provide an estimate of the frequency of such a crash occurring in the overall population for each given year. The data from both FARS and NASS-GES are derived from police accident reports.

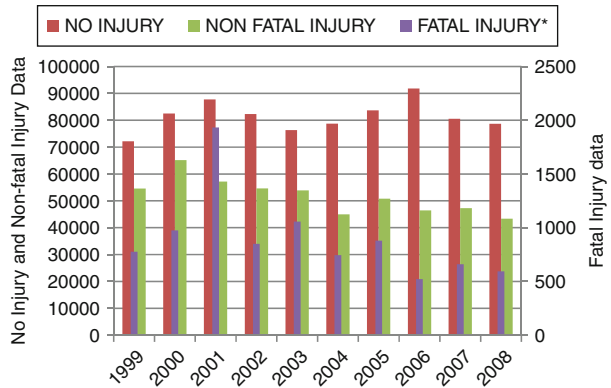
➤ *Figure 36.1* shows the number of fatal crashes (vehicles) and fatalities (people) due to drowsiness and fatigue over a 10-year period from FARS (Fatal Accident Reporting System) data. ➤ *Figures 36.2* and ➤ *36.3* show NASS/GES drowsiness related data including crashes, injuries, and fatalities for the period 1999–2008. Note that the number of fatalities is generally underestimated in the weighted NASS/GES data as compared to FARS data, which is the actual census.

In general, 2008 highway safety data reflects a reduction in the total number of injuries and fatalities, partially attributed to the reduced travel and miles driven due to the country’s economic downturn. Thus looking at prior data (i.e., 2007 and before), over 1,000 fatalities (from FARS), 47,000 injuries (from NASS/GES), and 80,000 crashes annually are caused by drowsiness and fatigue, indicating a significant safety problem. It is also believed that the statistical data underestimate the true magnitude of the problem because of difficulty and uncertainties in attributing the crashes to drowsiness and fatigue after the fact, i.e., during the accident data collection. The statistics definitely reflect the high risk of drowsiness driving and justifies the need for development of countermeasures.

2 Characteristics of Drowsiness and Fatigue

There have been several studies on driver drowsiness conditions. The following summarizes some of the important findings of a significant expert panel study by NIH – National Institute of Health and NHTSA – National Highway Traffic Safety Administration (Strohl et al. 1998) along with other references.

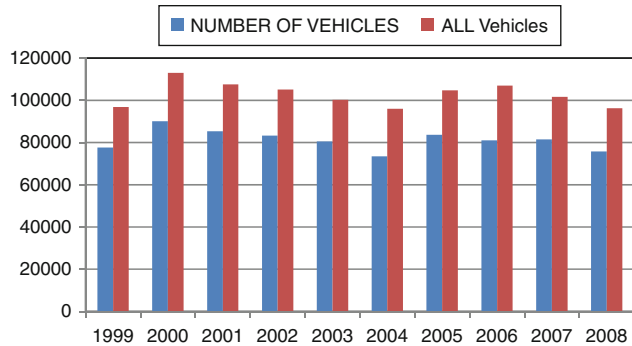
Year	GES Weighted Estimates (PEOPLE)		
	No Injury	Non-Fatal Injury	Fatal Injury*
1999	72161	54574	777
2000	82526	65186	976
2001	87753	57153	1933
2002	82333	54625	850
2003	76309	53899	1058
2004	78738	44986	745
2005	83664	50778	879
2006	91822	46460	522
2007	80571	47263	660
2008	78670	43349	594
Total	814547	518274	8994



■ Fig. 36.2

NASS/GES data of occupants non-injured, non-fatally injured, and fatally injured in crashes due to drowsiness in 1999–2008 period (Note: NASS/GES underestimates fatalities)

Year	Number Of Vehicles	All Vehicles
1999	77602	96778
2000	90025	112983
2001	85282	107516
2002	83220	105087
2003	80493	100213
2004	73443	95945
2005	83685	104712
2006	81023	106936
2007	81517	101672
2008	75760	96212
Total	812050	1028054



■ Fig. 36.3

Number of vehicles whereby the driver of the vehicle was reported as "Sleepy, or fell asleep" and total number of vehicles involved when at least one driver in the crash was reported as "Sleepy, or fell asleep"

Drowsiness or sleepiness is a condition of the human body which requires the person to sleep. It is a function of sleep/wake cycle of the body, governed by the two homeostatic and circadian factors. Homeostasis relates to the neurobiological need to sleep after long periods of wakefulness, making it hard to resist sleeping (Dinges 1995). "The circadian pacemaker is an internal body clock that completes a cycle approximately every 24 h.

Homeostatic factors govern circadian factors to regulate the timing of sleepiness and wakefulness.” (Strohl et al. 1998)

Fatigue is a condition of the body related to “weariness or exhaustion from labor, exertion, or stress” (Webster Dictionary); it is associated with a temporary loss of power and the inability to function properly or complete tasks. While drowsiness is a natural state of the body, fatigue is a condition or state, which is induced by external factors. Fatigue can cause sleepiness or vice versa. However, the commonality between the two is their effect on the loss of human sensory processing, perception, and the various functioning abilities of the driver. Quantification and hence identification of fatigue may be more difficult because of the variety and complexity of its symptoms. In general, there are better indicators of drowsiness on human body. Some conditions and causes of drowsiness are listed:

- Sleepiness can occur due to disturbances to the circadian sleep/wake cycles.
- Two sleepiness peaks are observed during the afternoon and before the night sleep time.
- Loss of one night sleep causes extreme short-term sleepiness.
- Habitual sleeplessness of one or two hours per night could cause chronic sleepiness during the day.
- Untreated sleep apnea syndrome (SAS) and narcolepsy could cause sleepiness.
- Sleepiness and performance impairment are neurobiological responses of the human brain to sleep deprivation and cannot be overcome by training, education, motivation, or other methods.
- Despite the intention or urge to stay awake, micro-sleeps may occur.
- Night workers, air crews, and travelers who cross several time zones, whose sleep is out of phase with the sleep/wake cycle, can experience sleep loss and sleep disruption which reduce alertness (Akerstedt and Kecklund 1994; Samel et al. 1995).

The following are construed consequences of drowsiness on the human abilities based on Strohl et al. (1998) workshop and drowsy driver experiments in driving simulators (Sayed et al. 2001 and Mortazavi et al. 2009):

- Drowsiness causes a delayed sensory processing ability and perception.
- Longer periods are needed to react to external stimulus during driving.
- Driver’s response and ability to control the vehicle degrades substantially.

In summary, under the drowsiness conditions the ability of the humans to operate and perform tasks degrades substantially. This causes a serious impairment during driving. Total loss of control occurs when drivers completely fall asleep at the wheel.

3 Characteristics of Drowsiness and Fatigue-Related Accidents

It is hard to determine exactly what percentage or what type of crashes occur due to drowsiness and fatigue. After crashes resulting in no harm, the drowsiness symptoms may

vanish due to the hyperarousal and anxiety of the drivers involved in the crash. Those crashes which result in fatalities rely on subsequent assumptions and deductions of the accident investigators and police reports. Only those data can be partially trusted that are from injury resulting crashes after which the drivers admit to drowsiness. The statistical data does not include this information, either. Despite these shortcomings, the available crash data and simulated experiments of drowsy drivers reveal certain common characteristics.

Horne and Reyner (1995) identified some of the vehicle- and environment-related criteria by which drowsiness-related vehicle crashes could be identified; these include:

- Vehicle running off the road
- No sign of braking
- No mechanical defect
- Good weather
- Elimination of speeding

If additional information about the driver (from surviving crashes) is also available, it would help in this identification process. Researchers have found many factors that may influence driver fatigue/drowsiness. Some of the important driver-related factors are:

1. Greater daytime sleepiness, more difficult schedules and hours of work, driver's age, driver experience, cumulative sleep debt, presence of a sleep disorder, and time of day of the accident (Gander 1999; McCartt et al. 2000)
2. Hours of continuous wakefulness before driving
3. Loss of sleep; duration of last sleep period (Carskadon and Dement 1981; Dinges 1995; Stutts et al. 2003; Mitler et al. 1997; Sweeney et al. 1995)
4. Time of the day, night time (Dinges 1995; Hertz 1988; Jovanis 1991; and Harris et al. 1972)
5. Sleep disorder (Stoohs et al. 1993; Young et al. 1997; Stutts et al. 2003)
6. Consumption of drowsiness causing medications
7. Monotony/length of driving (Akerstedt et al., 1994; McCartt et al. 1996; Fell 1994; Sagberg 1999; Thiffault and Bergeron 2003a)
8. Driver personality and age (Artaud et al. 1994; Thiffault and Bergeron 2003b; Campagne et al. 2004)

Sayed et al.'s driving simulator experimentation showed that drivers who after a normal night of sleep had 17–18 h of continuous wakefulness (a long day from 6 AM to 12 AM midnight) crashed within less than an hour of driving past midnight. Mortazavi et al. had a somewhat similar finding on experiments conducted with commercial vehicle drivers in a truck driving simulator. After 18 h of wakefulness, these subjects were asked to drive a monotonous roadway for over an hour in periods from 12 AM to 3 AM. Almost all subjects in this experiment also revealed symptoms of fatigue and drowsiness and crashed. Other important statistics on crashes attributed to drowsiness from earlier NHTSA data are:

- The highest numbers of crashes occur during the period from midnight to early morning.
- More than 40% occur between 1 AM and 7 AM.

- About 70% of crashes occur on rural highways with 55–65 mph speed limits. This generally provides a monotonous driving condition, prone to drowsiness.
- The first crash events are: 64% collisions with fixed objects (trees, guardrail, highway sign, etc.), 17% collisions with another moving vehicle, 7% are rollovers, and 6% are collisions with parked vehicles.

The severity of these crashes is typically high due to the significant loss of driver's control, which may result in either delayed or no braking response, leading to run-off-road, rollover, or a high-speed collision with other vehicles or barriers.

4 Drowsiness and Fatigue Detection Systems

Drowsiness detection methods can be divided into two major categories: (1) Detection by measuring and observing the driver physiological symptoms and conditions and (2) Detection by measuring the vehicle variables and states, which are caused by the control actions of the driver. The latter, obviously is still dependent on the drivers' condition and control action, but it does not require any direct measurement or monitoring of the driver. Each method has advantages and shortcomings.

Each method requires, first, finding strong correlation between drowsiness/fatigue and one or more corresponding detectable variables, regardless of whether it is a human-related or vehicle-related variable. Fortunately, prior research shows such strong correlations do exist and can be measured, albeit not easily, and in some instances not practical for real-life driving scenario. Second, the identified variable(s) that correlate(s) well with drowsiness need to be measured accurately and reliably. Then, in order to have a detection method, a hypothesis should be developed which determines or defines the drowsiness condition. The variables' patterns or thresholds based on the hypothesis should be defined through experiments with a statistically significant population in simulator, track, or field operational tests. Finally, a detection method can be developed based on the measurements (via sensors), the hypothesis which defines a classification schema, and/or threshold values to indicate a possible drowsiness state of the driver. The detection method combined with a warning/alert or automatic vehicle control serves as a driver assistance safety system to mitigate drowsiness driving.

5 Sensing Driver's Physical and Physiological Conditions

Eye closure: Monitoring driver eyes is one of the most successful techniques for detecting drowsiness and is studied by many researchers. Dingus (Dingus et al. 1985) developed the PERCLOSE algorithm, which is a measure of the proportion of time that the eyes of a subject are closed over a certain time period. Different techniques have been used to track the eyelid closures: Torsvall and Akerstedt (1988) used Electrooculography (EOG) to detect eye movements; Ogawa and Shimotani (1997) used the angle of inclination of eye corners to track the eyelid closures. Seki, Shimotani, and Nashida, (1998) used reflection

from the retina for capturing eye closure. Eye closure activity can provide good detection accuracy but changes in light conditions, correction glasses, angle of face, and other conditions can seriously affect the performance of image processing systems.

Electroencephalogram (EEG): EEG recorded from the human scalp is the most important physiological indicator of the central nervous system activation and alertness. In time domain, commonly used EEG measures include average value, standard deviation, and sum of squares of EEG amplitude, while in frequency domain energy content of each band (β , α , θ , δ), mean frequency, and center of gravity of the EEG spectrum are commonly used (Huang et al. 1996, Lal and Craig 2001, Wu et al. 2004).

Facial expressions and body posture: According to Wierwille and Ellsworth (1994) trained observers could rate the drowsiness level of drivers based on video images of driver faces. Bergasa et al. (2003) and others have developed vision-based systems to extract facial expressions automatically. There is little evidence about the accuracy and robustness of such systems. Another chapter is dedicated to advances and successes of vision systems for facial monitoring. Andreeva et al. (2004) studied changes in body postures by attaching triple-axis accelerometers to upper body but the results reported are inconclusive.

Other physiological conditions have also been monitored that include changes in heart beat, blood pressure, skin electrostatic potential, and body temperature but all with limited success.

6 Sensing Vehicle State Variables

Drowsiness causes anomalies in driver control actions which results in notable changes in vehicle states (steering/lane keeping, accelerating, etc.). Methods have been developed by measuring and detecting the distinguishing features of different states. These methods have the advantage of being nonintrusive to the drivers. The focus of these secondary task measurements is not on the condition of driver but on the output of the vehicle as controlled by the driver. Vehicle variables that can be measured include vehicle speed, acceleration, yaw rate, steering angle, and lateral displacement. In this section, first the adverse effects of drowsiness on vehicle state variables (identifying variables and defining hypothesis) are identified and later on different variables that are proposed by researchers as performance measures that can potentially be used in a drowsiness detection system are introduced. Finally, different real-time detection systems which leverage vehicle state variables to detect drowsiness are discussed.

7 Correlation between Drowsiness and Vehicle State Variables

Wierwille et al. (1992) discussed the performance measures as indicators of driver drowsiness in detail. In this section, these measures are classified into four groups:

- Vehicle Steering Activity
- Vehicle lateral position

- Vehicle speed
- Yaw/Brake/Acceleration

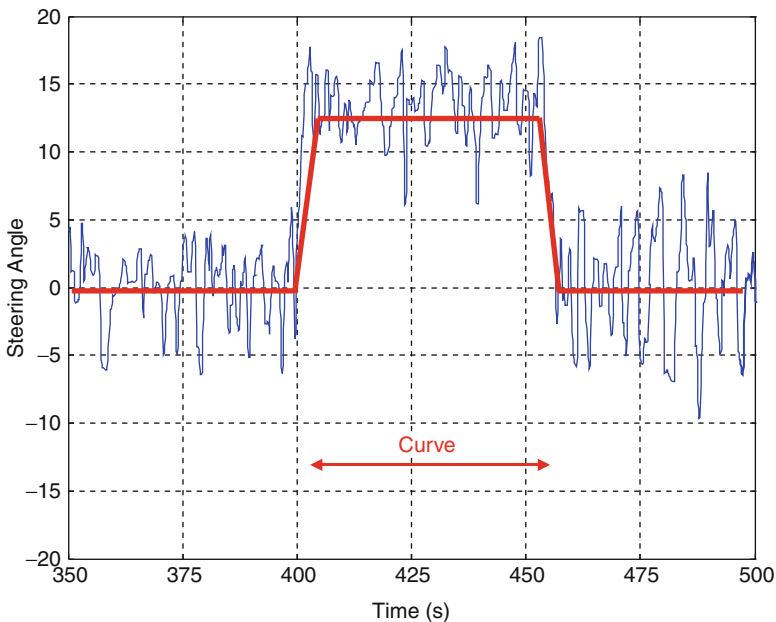
The following is a summary of these measures.

7.1 Vehicle Steering Activity

Vehicle steering has been cited by many studies as a characteristic variable which can predict the driver drowsiness. Selected studies are further explained below.

While driving, drivers make two types of steering adjustments/corrections to control a vehicle (see [Fig. 36.4](#)):

- Micro-corrections: These adjustments can be traced as small amplitude oscillations in steering wheel angle plots which keep a vehicle in the center of a lane (lane keeping adjustments).
- Macro-corrections: In addition to lane keeping adjustments, drivers may make large steering adjustment to negotiate a curve or change lanes. The corresponding steering wheel angle signal looks like a riding wave in this scenario.



■ Fig. 36.4

Micro- and macro-adjustments in a steering wheel angle waveform while a curve negotiation scenario

Researches show that sleep deprivation adversely affects the two aforementioned steering adjustments in forms of larger amplitude steering corrections (over-steering) and less frequent micro-corrections (Sayed et al. 2001; Mortazavi et al. 2009).

Based on the results of a truck simulation experiment, Mortazavi et al. 2009 identified two major steering control degradation phases due to drowsiness. In the first so-called “impaired” phase (Phase-I), driver’s decision-making ability is affected, and the result is a zigzag style driving in which the driver cannot smoothly follow the desired trajectory. During zigzag driving, standard deviation of vehicle lateral displacement and steering wheel angle increases as compared to normal driving (over-correction of steering). This phenomenon occurs on and off for a period of time before the driver dozes off. In the second so-called dozing-off phase (Phase-II), the driver provides no corrective feedback to control the steering wheel and the vehicle continues its path without any correction. This could be traced by a constant steering angle value over a short period of time (smaller variability in steering wheel angle) combined with increasing lateral displacement (lane departure). ● *Figure 36.5* highlights a sample of described phases for a drowsy driver, and compares it with an alert driver’s steering wheel signal sample. The figure compares the steering and lateral displacement data before a crash to the alert data for the same segment of the road.

Here is a list of steering-related performance measures – proposed by different studies – that have direct correlation to drowsiness:

7.1.1 Frequency of Steering Wheel

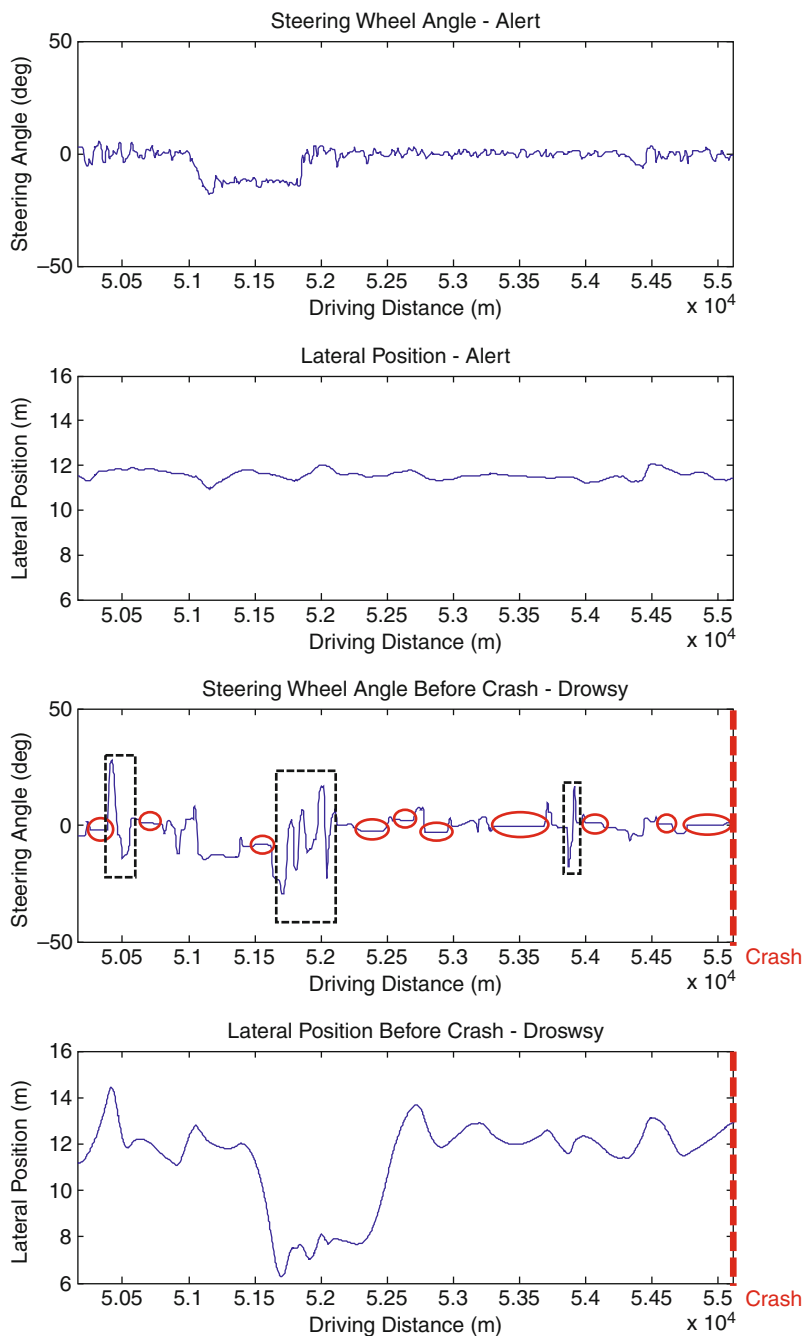
Any steering wheel pass across zero degree is counted as a reversal. Sleep-deprived drivers have lower frequency of steering reversals Reversal (Hulbert 1972; Ryder et al. 1981).

7.1.2 Steering Wheel Reversal Rate

Drowsiness decreases steering wheel reversal rate (reversal per minute) with 1.0° gap size – angle that the steering wheel must be reversed before being counted as a reversal (Elling et al. 1994). Also, number of large amplitude steering movement (exceeded 15° after steering velocity passed through zero) and medium amplitude movement (exceeded 5° , but did not exceed 15° , after steering velocity passed through zero) increases with level of drowsiness.

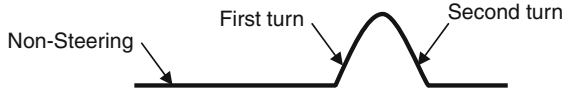
7.1.3 Steering Correction

This study hypothesizes that when a driver is drowsy or falling asleep his/her steering behavior becomes more erratic, that is, “more frequent steering maneuvers during



■ Fig. 36.5

Examples of dozing off intervals highlighted on a steering wheel angle signal of a drowsy driver compared with the signal plotted (from same stretch of the road) while the driver was alert (*rectangular*: impaired phase; *ellipses*: dozing-off phase)



■ Fig. 36.6

Typical steering pattern

wakeful periods, and no steering correction for a prolonged period of time followed by a jerky motion during drowsy periods (Yabuta et al. 1985)". Yabuta et al. assume that a typical steering wave-form consists of a non-steering period followed by two steering corrections in opposite directions – so-called “first turn” and “second turn” respectively (see ► Fig. 36.6).

The study found out that drowsiness affects three parameters:

1. Occurrence of frequency of large non-steering periods when magnitude of the first and second turns are fixed to 3° or more
2. Occurrence of frequency of large steering wheel turns when magnitude of the non-steering period is greater or equal to 1 s
3. Occurrence of frequency of the first turn speed when non-steering period as well as the first turn and the second turn magnitudes are fixed (≥ 1 s, $\geq 6^\circ$ and $\geq 3^\circ$ respectively)

7.1.4 Steering Velocity

Drowsiness and sleep deprivation decreases steering velocity and increases standard deviation of steering velocity (Dingus et al. 1985; Elling et al. 1994).

7.1.5 Steering Amplitude Duration Squared Theta

This parameter shows significant correlation with the subjective evaluation of drowsiness and is defined as follows in ► Eq. 36.1 (Siegmund et al. 1996):

$$Am_{D2\Theta} = \frac{100}{N} \sum_{j=1}^J \left(A_j^\theta t_j^\theta \right) \quad (36.1)$$

where A_j^θ is the j th block in the $(\theta - \theta_m)$ data of a leg (64-s time period)

θ_m = mean steering angle for a leg being analyzed

t_j^θ = the length of the j th area block in the leg limited between two consecutive zero-crossings;

J = the total number of area block in the leg

N = total number of samples in leg, and 100 is a scaling factor

Sampling frequency = 128 Hz

7.1.6 Weight Flat Zero

This study shows phase plots of steering angle versus steering wheel velocity (θ_i, ω_i) can be used as an indicator of drowsiness, if only points satisfying the condition $|\omega| \leq \omega_c$ with $\omega_c = 5^\circ/\text{s}$ are included in the calculation (Siegmund et al. 1996). All points satisfying this condition are weighted by the square of the distance from the origin. Weight Flat Zero is defined as follows in Eq. 36.2:

$$Wt \text{ Flat } 0 = \frac{100}{N} \sum_{i=1}^N \left(\frac{(\theta_i - \theta_m)^2}{a^2} + \frac{\omega_i^2}{b^2} \right) \quad (36.2)$$

where θ_i = i th value of the steering angle

θ_m = mean steering angle for a leg being analyzed

a = half axis length of ellipse in θ dimension

b = half axis length of ellipse in ω dimension

ω_i = i th value of steering angular velocity

ω_c = cutoff value of omega ($5^\circ/\text{s}$)

N = total number of samples in leg, and 100 is a scaling factor

Sampling frequency = 128 Hz

7.2 Lateral Position

Maintaining Vehicle Lateral Position (i.e., lane tracking ability) decreases as the time on task increases (Mast et al. 1966). Skipper et al. (1984) found that measures related to vehicle lane position could be used to detect drowsiness. Variables such as the number of lane deviations, the standard deviation of lane position, and the maximum lane deviation are found to be highly correlated with eye closures. According to Dingus et al. 1985, the mean square of lane deviation and mean square of high-pass lateral position show good potential as drowsiness indicators.

Stein 1995 studied the effect of impairment on driving performance in truck drivers. Using data from a simulator experiment, Stein found that the standard deviation of lane position increases remarkably after the driver gets fatigued at 13 h of driving. The standard deviation of the heading error also began to increase after 13 h.

Pilutti and Ulsoy (1997) performed experiments on the Ford driving simulator for detecting driver fatigue. Their results demonstrate that only the standard deviation of lateral position shows significant change and correspond well with PERCLOS model.

7.3 Vehicle Speed

Vehicle Speed variability in general has not shown any significant correlation to drowsiness (Mortazavi et al. 2009). Safford and Rockwell (1967) reported no increase in speed

variability during a 24-h driving experiment. Riemersma et al. (1977) recorded vehicle speed during an 8-h night driving experiment and found an increase in the standard deviation of speed, calculated over 45-min intervals, after the first 3 h of driving. Mackie and O'Hanlon (1977) recorded speed in a 6-h driving experiment, with a 45-min pause after 3 h of driving; they found a regular increase in the standard deviation of speed from the third driving hour.

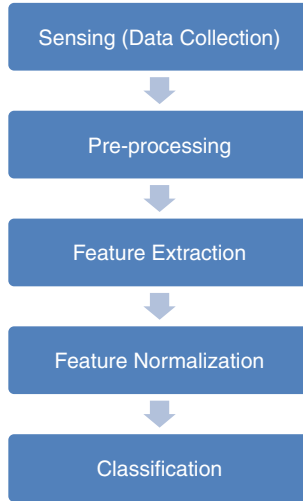
7.4 Yaw/Brake/Acceleration

Yaw/Brake/Acceleration activity or variances may have some relationship with drowsiness but are not proven to be solid indicators. Dingus et al. (1985) found that the yaw deviation variance and the mean yaw deviation (calculated over a 3-min period) show some promise to be considered as drowsiness indicators. However, several researchers have found no relation between drowsiness and vehicle yaw, brake, or acceleration. Safford and Rockwell (1967) analyzed data from a 24-h driving experiment and reported that the accelerator pedal reversals are correlated with driving time. But according to Dingus et al. (1985), there is little evidence of any relation between accelerator activity and time or drowsiness. Similarly, Brown 1966 found no evidence of any correlation between accelerator and drowsiness.

8 Drowsiness Detection Methods Based on Vehicle State Variables

There are two major detection schemas to address drowsiness detection problem. The first approach is to directly input vehicle state variables into a predefined mathematical model to estimate a parameter that can be directly correlated to the level of drowsiness. In other words, the level of drowsiness is predicted based on the output values of the model, for example, F4e-3 algorithm (See 🔍 Sect. 36.9).

The second approach uses pattern classification solutions to tackle the detection problem in which the goal is to classify the level of drowsiness into two different possible states/classes: alert or drowsy. 🔍 Figure 36.7 shows the processes of a typical pattern classification schema. The raw vehicle state data, measured by various sensors (or a single sensor), are preprocessed (i.e., filtering) and then passed to a feature extractor module, whose goal is to reduce the data by measuring specific features and properties characterizing a state of nature, that is, drowsiness. The feature extractor extracts specific parameters/evidences from the vehicle state data which represent effects of drowsiness on driver's steering control behavior. In the next step, the features would be normalized (if needed). Then, a classifier uses the normalized evidences to decide the true state of nature (i.e., level of drowsiness).



■ Fig. 36.7

Typical pattern classification schema used in drowsiness detection domain

9 Examples of Drowsiness Detection Methods

In this section, four drowsiness detection methods are presented.

Using backward stepwise multiple linear regression, Wirewille et al. 1994 developed a simple algorithm (F4e-3) that estimates PERCLOS (ePERCLOS):

$$\begin{aligned}
 ePERCLOS = & -0.00304 + 0.000055(STVELV) - 0.00153(LGREV) - 0.00038(MDRVE) \\
 & + 0.003326(LNMNSQ) + 0.00524(LANVAR) - 0.00796(INTACDEV)
 \end{aligned}
 \quad (36.3)$$

In which:

- *STELV*: Variance of steering wheel velocity (degrees/sec).
 - *LGREV*: The number of times that the steering wheel movement exceeded 15° after steering velocity passed through zero.
 - *MDREV*: The number of times that the steering wheel movement exceeded 5°, but did not exceed 15°, after steering velocity passed through zero.
 - *LNMNSQ*: The mean square of lane position with respect to lane center measured in feet.
 - *LANVAR*: The variance of lateral position relative lane center, in feet.
- INTACDEV*: The standard deviation of the lateral velocity of the vehicle, measured in volts where 1 volt equals 73.34 ft/s.

Fukuda et al. (1995) developed a driver drowsiness detection system at the Toyota Motor Company that used steering adjustment time to estimate drowsiness. This method consists of the following steps:

1. Steering adjustment intervals are calculated at different speeds for alert conditions (learning phase). These intervals vary with speed and individual behavior but it follows the same pattern. Fukuda et al. specified four adjustment patterns. Their analysis of vehicle steering data showed that: (1) minimum adjustment steering angle is 0.5° and (2) minimum adjustment steering interval is 0.5 s. For a pattern, given $\delta(n)$ is a steering point sample, the steering adjustment interval is defined as the period in which n satisfies the following criteria:

$$\begin{aligned}\delta(n) &\geq \delta(n-k) \\ \delta(n) - a_n \times \frac{k}{6} &\geq \delta(n+k) \\ (k &= 1 \text{ to } 6), a_n = 0.5\end{aligned}\quad (36.4)$$

Sampling resolution and frequency is 0.1° and 128 ms respectively.

2. The steering adjustment intervals are normalized at 80 km/h speed. These intervals are constantly calculated. Whenever it reaches a threshold value, the driver is classified as drowsy. The value of drowsiness threshold (D_e) is not constant but it varies with speed. During alert state, the mean value of learned steering adjustment intervals is calculated. The driving threshold is estimated by taking the product of the mean value of learned steering adjustment intervals in the normal state, coefficient of correction, and coefficient of drowsiness judgment. Therefore, the system classifies a driver as drowsy when:

$$D_e > G_I(D_r) \times G_v(V) \times D_r \quad (36.5)$$

where $D(i)$ = steering adjustment interval

D_r = mean value of learned steering adjustment

D_e = mean value of most recent steering adjustment intervals

$G_v(V)$ = coefficient of correction against the fluctuation of steering adjustment interval in the normal state according to the variation of vehicle speed

$G_I(D_r)$ = coefficient of drowsiness judgment according to the absolute values of steering adjustment intervals of each driver

$$D_r = \frac{\sum_{i=1}^n D(i)}{n} \quad (36.6)$$

$$D_e = \frac{\sum_{i=m-p}^m D(i)}{p} \quad (36.7)$$

m = most recent interval, $m < n$

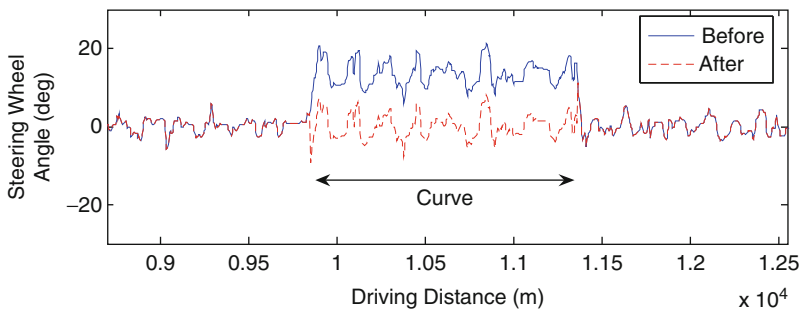
$p = 50$

The results showed good correlation with EEG.

Sayed and Eskandarian (2001) developed an unobtrusive method of detecting drowsy drivers by monitoring the steering activity of drivers using a trainable artificial neural network. An Artificial Neural Network (ANN) is trained to learn steering input of drivers under different levels of drowsiness (alert and drowsy). This system was also used to detect truck driver drowsiness based on steering activity (Eskandarian and Mortazavi 2007). The training of the neural network was based on the learning of the phase-I steering performance degradation (see ► Sect. 36.7). This method is based on assuming the adverse effect of drowsiness on steering macro-corrections. The data of the experiment on passenger and truck drivers in a simulated environment was used to train and test the ANN.

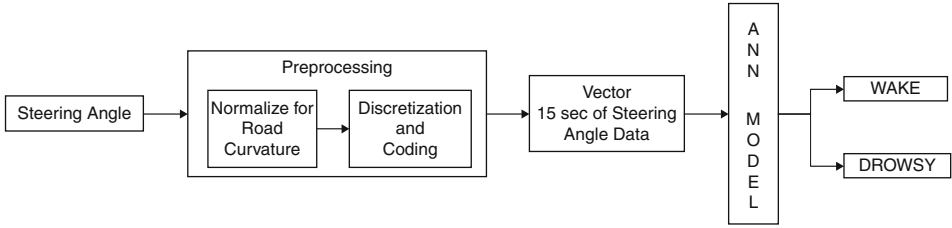
Prior to training the network, the effect of road curvature on steering wheel angle is removed (preprocessing). Road horizontal geometry normally includes two types of geometric sections, that is, straight lines and curves. In the straight sections, the steering angle signal consists of only the steering adjustments for lane keeping. In the curve sections, including clothoids, which connect straight and curve segments, the steering signal contains the waveform for lane keeping as well as the waveform for negotiating road curvature. Sayed and Eskandarian (2001) proposed that the effect of curvature could be removed from steering wheel angle signal by subtracting the signal trend from the original signal. They used a simplified and modified procedure of trend extraction. In this method, if four or more consecutive data points are of the same sign (positive for right turn and negative is for left turn) and their sum are greater than or equal to 15° (absolute), then all these points are assumed to be from a curve or portion of a curve section. The mean value of these data points is then subtracted from each of these points (see ► Fig. 36.8).

► Figure 36.9 shows the schematics of the proposed ANN method. The processed steering data are also coded into a vector to be applied as input into the network (feature extraction). The sum of vectors on every 15-s interval is used as the ANN input. Two separate sets of preprocessed vectors were used to train and test the neural network (classifier).



■ Fig. 36.8

Steering wheel data before and after road curvature removal



■ Fig. 36.9

Schematics of the ANN method

For the input of the ANN, the steering signal was discretized and coded into a 1 by 8 vector. The steering angle amplitude is divided into eight smaller ranges, r_1 to r_k . These ranges are defined as follows:

$$-\sum_{k=1}^4 p_k > r_i > \sum_{k=j-1}^4 p_k; i = (1, 2, 3, 4) \quad (36.8)$$

$$\sum_{k=9-i}^4 p_k > r_i > \sum_{k=8-i}^4 p_k; i = (5, 6, 7, 8) \quad (36.9)$$

where p_k are constant. By choosing different values for p_k , the coded vector can be calibrated for different driving behaviors. Some drivers make small and accurate steering correction (low amplitude), while others are less sensitive to their lane keeping and make larger steering movement (higher amplitude) in their normal driving behavior. Larger values for p_k are used for drivers with large steering movement to make discretization ranges wider. p_0 and p_4 represent upper and lower steering angle limits respectively ($p_0 = 900$ and $p_4 = 0$). Over a given period of time, T , if the mean steering value fell into one of the ranges represented by r_i , the i th component of the eight-dimensional vector state, $I(T)$, was set to 1. The other indices values are equal to zero. For this study T is one second.

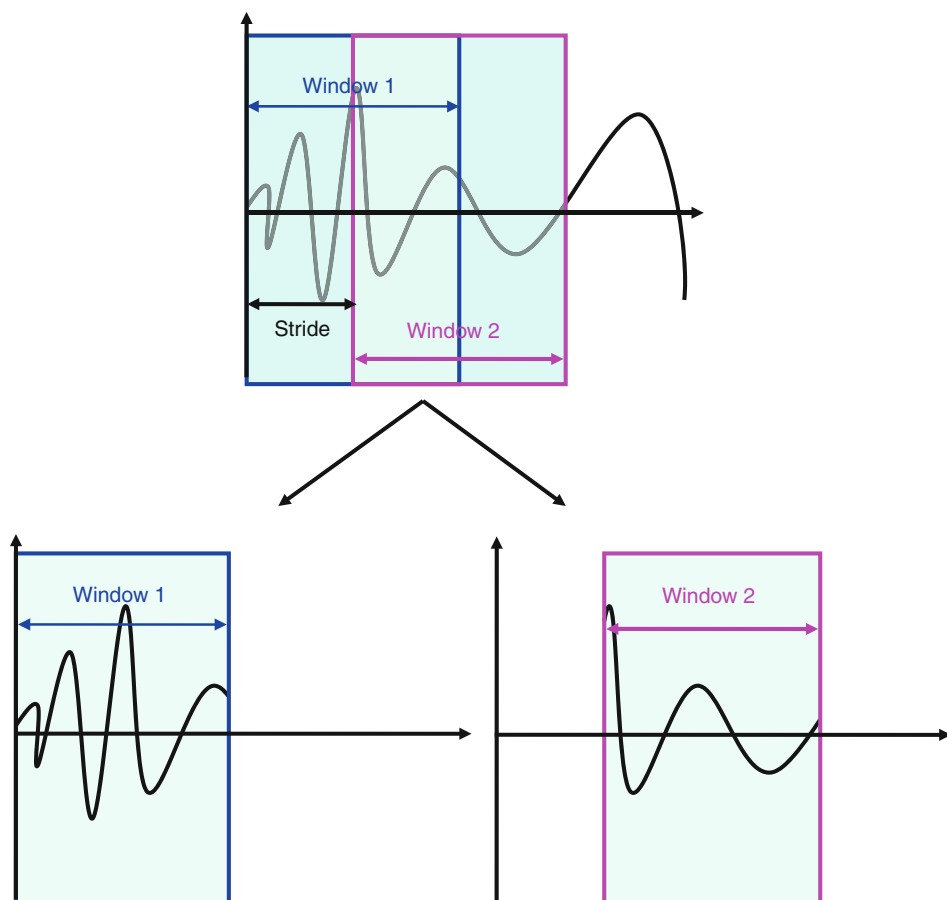
After vectorizing the mean steering for each second, each vector was summed over an interval of n point resulting in ANN input vector $X(n)$:

$$X(n) = I(T) + I(2T) + \cdots + I(nT) \quad (36.10)$$

For $n = 15$, $X(n)$ represents 15 s of steering activity.

During the supervised training of the neural network, the input vectors, $X(n)$, were classified into two output vectors. Vector $[1,0]$ represents alert state and vector $[0,1]$ demonstrates drowsy state.

The result of the tests in a passenger vehicle driving simulator demonstrated over 85% success rate in detection accuracy and less than 14% false alarms. The truck experiment



■ Fig. 36.10
Signal windowing

also showed similar results. The shortcoming of this method is the need for a large set of exemplars for training the neural networks which requires samples (steering time histories) of both alert and drowsy driving.

Eskandarian and Mortazavi (2009) provisionally patented a new method of detection based on a signal processing scheme on steering signal. In contrary to other methods, they looked into variation of steering wheel angle along with driving distance (vs s waveform). The method was developed and validated based on data collected from commercially licensed drivers tested under alert and sleep-deprived conditions in a truck driving simulation environment (Mortazavi et al. 2009). Consistent with ● Fig. 36.7, the proposed method comprises of five steps:

Sensing (Data Collection). Data is collected and analyzed consequently over small overlapped frames or windows. ● Figure 36.10 shows schematics of data windowing. The objective is to extract and process specific features from each steering data frame.

The distance between the beginnings of two consequent windows is “stride.” In their study, Mortazavi and Eskandarian proposed window size of 1,000 m and stride size of 250 m.

Preprocessing. The sampling frequency of measured data may not be uniform. As the first step, the data are down-sampled. Proposed sampling rate is 10 Hz for this study.

Feature Extraction. The method implements a signal decomposition method, Empirical Mode Decomposition (Huang et al. 1996), to decompose steering signals into different modes. The advantage of implementing the Empirical Mode Decomposition (EMD) method is that the extraction procedure does not require any preprocessing for elimination of road curvature effect from the steering signal. The detection method looks for anomalous signal behaviors in terms of amplitude and frequency in a specific extracted signal (mode) to detect drowsiness.

Empirical Mode Decomposition is an empirically based data analysis method, developed by Huang et al. (1996) to handle non-stationary and nonlinear signals. EMD method decomposes a signal into a posteriori-defined basis, and is an adaptive method derived from the data. The basic assumption is that any data comprises different intrinsic modes of oscillation. Technically speaking, this method is a way of representing a signal in terms of Amplitude Modulation (AM) and Frequency Modulation (FM). Each intrinsic mode is representing a simple oscillation with equal number of extrema and zero-crossings.

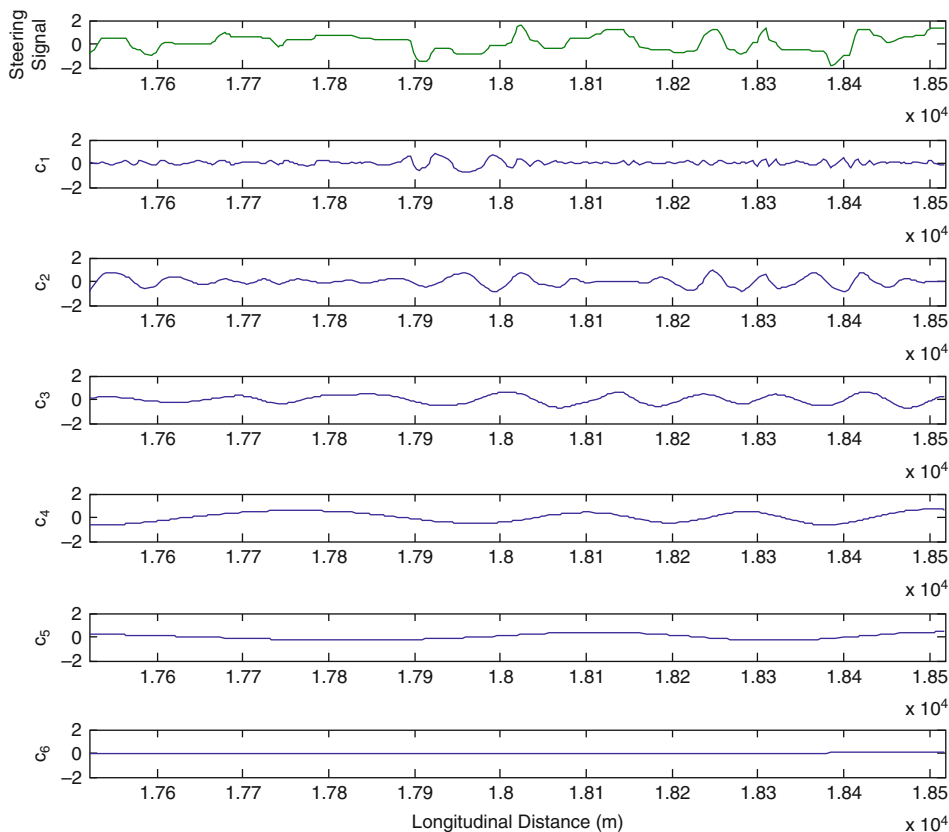
The goal of decomposition is to represent a signal in terms of different intrinsic modes with simple oscillation. Each mode, so-called Intrinsic Mode Function (IMF), has the same number of maxima and zero-crossings. (They can differ at most by one.) In addition, at any point of an IMF signal, the mean value of the envelope passing through maxima (upper envelope) and the envelope passing through minima (lower envelope) is zero (Huang et al. 1996; Flandrin and Goncalves 2004). The process of extraction IMF functions is called sifting. Given a signal $x(t)$, the sifting process algorithm output will be a series of c_j (IMFs) and a residual signal (r_n) where:

$$x(t) = \sum_{j=1}^{n-1} c_j + r_n \quad (36.11)$$

► *Figure 36.11* shows a sample of decomposed steering wheel angle signal (for more technical information on EMD see Huang et al. 1996; Flandrin and Goncalves 2004; Rilling et al. 2003; Huang and Shen 2005).

The core objective of this detection method is to capture the previously introduced two degradation phases, Phase I and Phase II (i.e., impaired and doze off phases). Two features are extracted from the first IMF signal (IMF_1 or c_1) which represent Phase I and Phase II behavior. The study shows that drowsiness has two adverse effects on IMF_1 signal when compared to signals extracted from a normal driving condition. Each effect is directly correlated to one of the two degradation phases:

1. Phase I. During Phase I, the steering signal has larger amplitude steering corrections (over-corrections). Consequently, a similar effect on IMF_1 signals can be detected.
2. Phase II. The constant value interval of the steering wheel signal during dozing off periods can be inferred as the intervals when the local frequency is zero. As a result, during Phase II,



■ Fig. 36.11

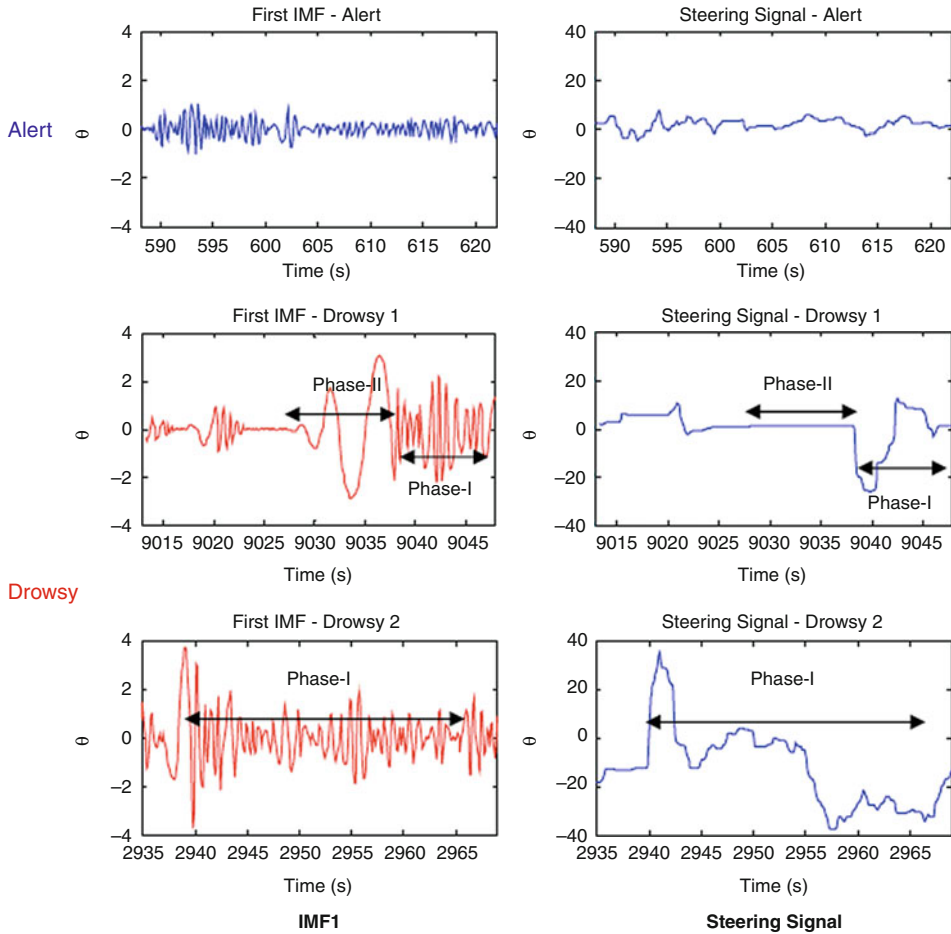
A sample of decomposed steering signal using EMD

distances of zero-crossings are longer. This characteristic was detected as a relative large-scale oscillation on IMF_1 signals. Besides, longer distances of zero-crossings were expected for the intervals with slower corrections (low local frequencies).

► [Figure 36.12](#) shows examples of IMF_1 signals affected by different phases of drowsiness degradation and compares them with an alert signal. The effect of each phase is also marked on the IMF_1 plots.

Mortazavi and Eskandarian (2008) proposed two parameters to capture Phase I and Phase II drowsiness-related incidents:

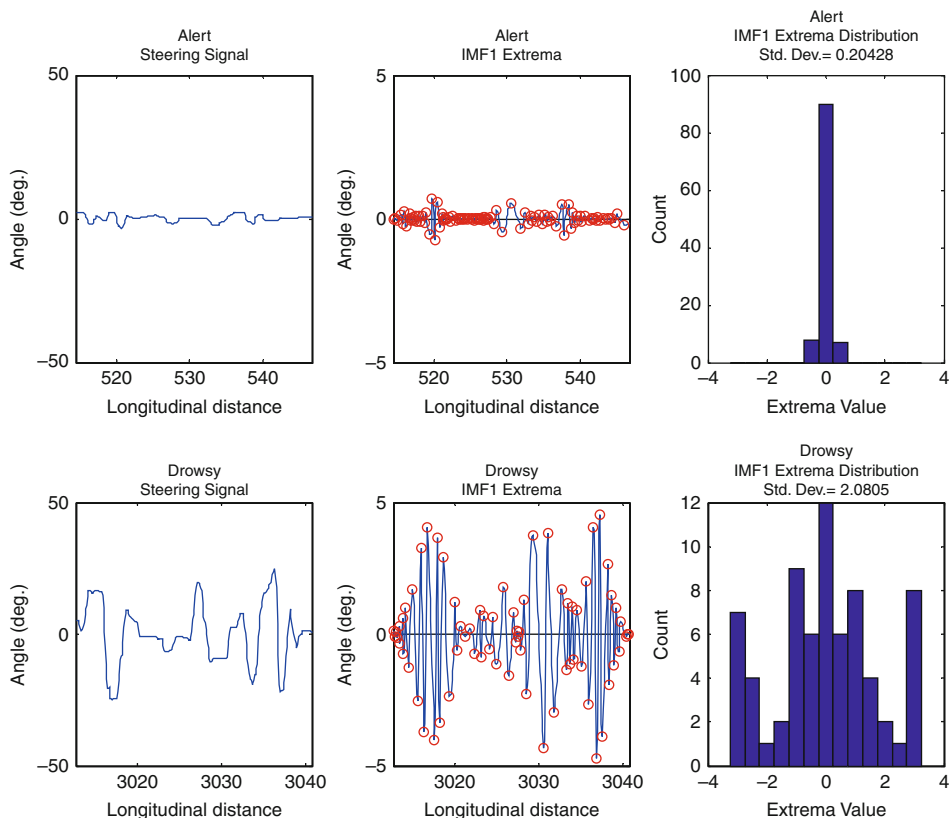
1. *Feature 1 – Standard Deviation of IMF_1 Extrema Distribution (SDIE)*. This feature measures Phase-I steering control degradation. The goal was to extract a feature from the first IMF signal that quantified the over-steering behavior of a drowsy driver. When the driver is drowsy, the distribution of the IMF_1 extrema values – as well as the standard deviation of the distribution – reveals the fact that standard deviation of



■ Fig. 36.12
Examples of the effect of drowsiness on IMF1

IMF_1 extrema (SDIE) is generally greater than the SDIE values extracted from normal driving intervals (► Fig. 36.13).

2. **Feature 2 – Standard Deviation of IMF_1 Distance of Zero-Crossings (SDZC).** During the dozing off periods, large distance of zero-crossings (DZC) characterizes the effect of Phase-II steering control degradation phenomenon in IMF_1 signals, comparing to DZC values extracted from normal driving data. As described earlier, DZC also represents the signal local frequency. The standard deviation of zero-crossing distances measured in a data window is generally higher than the corresponding values for alert driving states. ► Figure 36.14 shows the distributions and standard deviation values for two alert and drowsy samples.

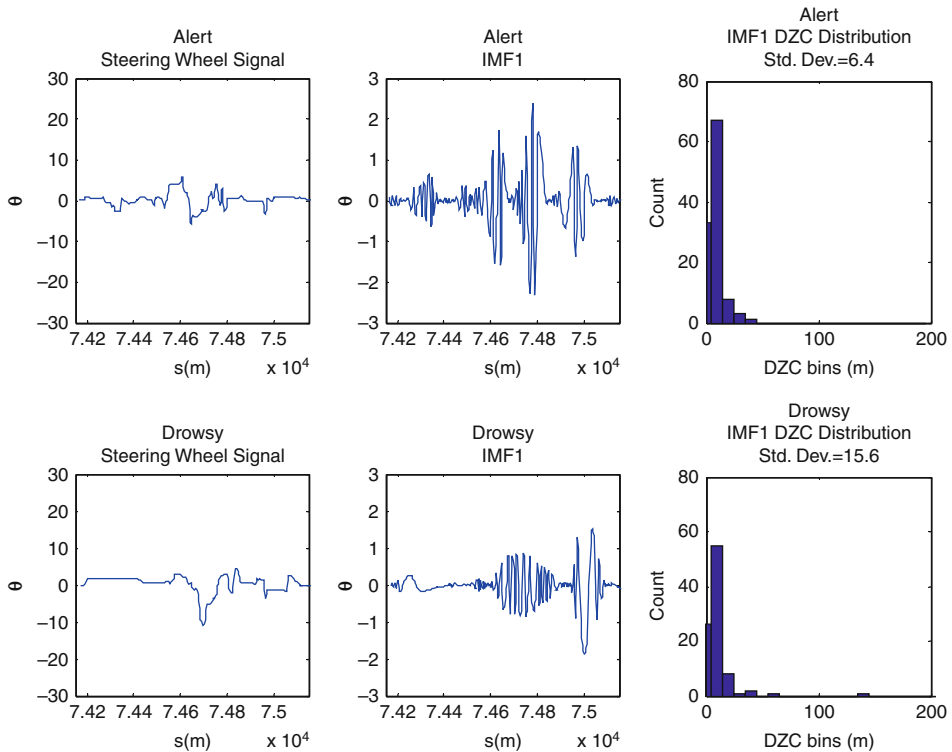


■ Fig. 36.13

Phase I effect on extrema distribution of IMF1

Post-processing. From each windowed signal the two features are extracted and paired to form a two-dimensional feature vector, ($SDIE$, $SDZC$). In this step, extracted feature vectors are smoothened and normalized.

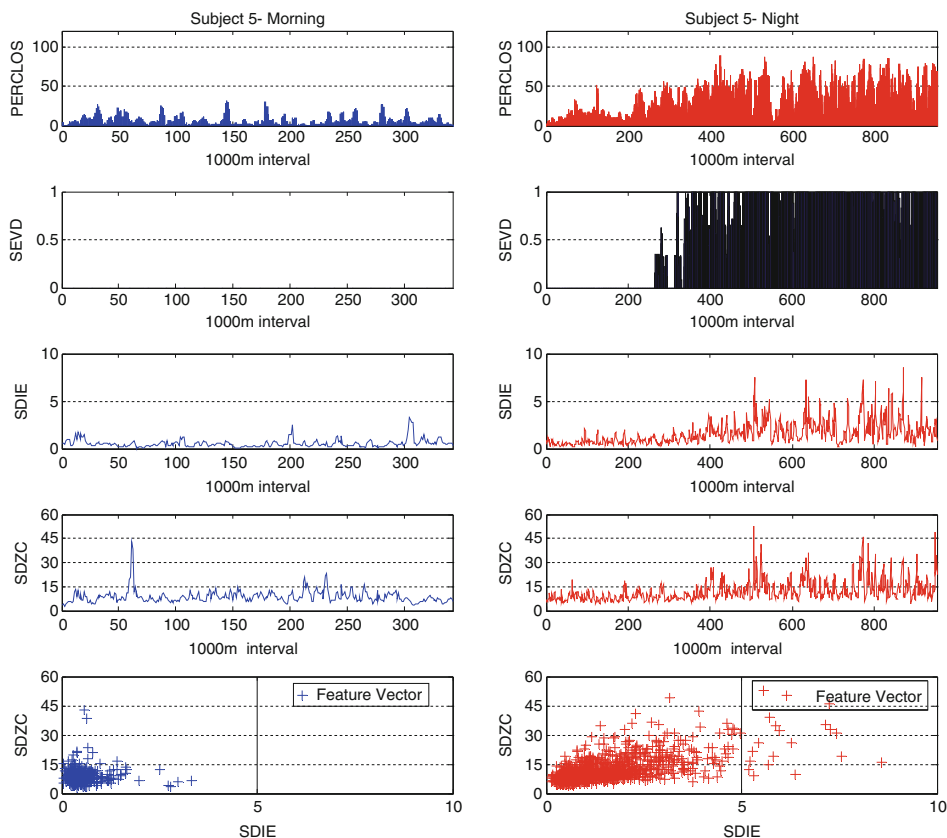
A single measurement from a window cannot represent the behavior of a signal. Instead, more observations and measurements of the descendant neighboring windows are required to have a better understanding of the modeled signal behavior. Averaging is the simplest characteristic that quantitatively models the data behavior. Therefore, the measured features were averaged over a constant number of consecutive windows for further analysis. The process is referred to as “ n -point averaging”. ▶ Figure 36.15 displays a sample of $SDIE$ and $SDZC$ values measured for each window ($Window\ size = 1,000\ m$; $Stride = 250\ m$) for a subject tested in the morning (alert) and at night (sleep-deprived) as well as the corresponding measured feature vector clusters. These graphs are also compared with PERCLOS and subjective evaluation of drowsiness (SEVD) where $SEVD = 0$ is alert and $SEVD = 1$ is extremely drowsy (Mortazavi et al. 2007).



■ Fig. 36.14

Samples of IMF1 distances of zero-crossings (DZC) distribution

The study of the feature vector clusters showed that – for alert driving state – the clusters' shapes and mean vectors differed among drivers. The difference was because of the unique steering control style of each driver. Some drivers prefer to control their vehicles with larger steering corrections – relatively higher $SDIE$ – while others tend to perform less frequent steering corrections – relatively higher $SDZC$ – during their normal driving behavior. The steering behavior can also include both mentioned correction characteristics. Theoretically, the shape of the cluster can be presumed as an ellipse with a long axis toward the direction of a feature that represents driver's dominant steering control behavior. As a result, each feature's normal range (extracted from normal driving), the cluster shape, and principle axes directions differ from driver to driver. Therefore, the feature vectors have to be normalized to remove the variations resulting from individual driving normal habits. The mathematical solution for this challenge lies in the concept of whitening transform (Duda et al. 2001). The extracted feature vectors from a normal driving sample data set generally create an oval-shaped cluster. The cluster directions of principle axes and the mean vector are different among drivers. This transformation will



■ Fig. 36.15

Comparison of SDZC and SDIE values extracted from steering wheel angle data sets collected from two separated simulation runs: morning run (alert) and night run (sleep deprived)

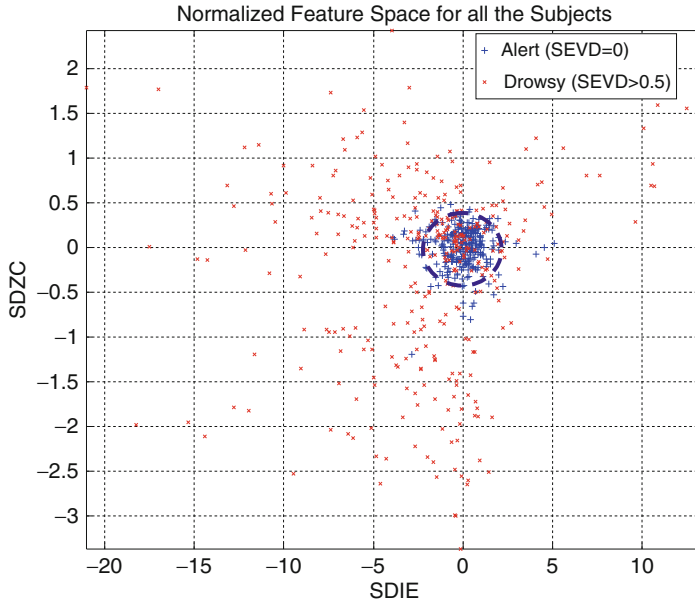
generate a unique shape for all clusters to be compared. In this process, a sample data set from alert driving state is collected and square matrix of A_w is calculated and multiplied by extracted feature vectors. The definition of A_w is:

$$A_w = \Phi \times \Lambda^{-1/2} \quad (36.12)$$

where Φ : a matrix whose columns are orthonormal eigenvectors of Σ (covariance matrix of new feature vectors)

Λ : diagonal matrix of the corresponding eigenvalues

Classification. ● Fig. 36.16 shows the concatenation of all the feature vector samples after normalization, collected from a truck simulation experiment. The accumulation of alert points around the origin of the coordinate system is obvious. The normalized alert feature vectors' cluster can be separated by a "discriminant circle." Ideally, the points inside the circle belong to the alert class and the points outside the circle fall into the



■ Fig. 36.16
Concatenation of feature vectors

drowsy category. However, there are few exceptions in which “drowsy” feature vectors are located inside the alert discriminant circle or vice versa. This is the result of the existence of occasions when the driver is drowsy but the steering control looks normal (missed detection), or the alert driver is forced to conduct an abnormal steering control because of the traffic maneuver (false detection). Mortazavi and Eskandarian tested the feature points using k-Nearest Neighbor classification method. This new method showed a similar success rate in detection as the earlier ANN method. The study also showed that the proposed algorithm was capable of issuing at least one warning before 97% of drowsiness-related lane departures.

10 Advantages and Disadvantages of Detection Systems Using Vehicle State Variables

Main advantages and disadvantages of methods using vehicle state variables compared to methods using driver’s physical and physiological conditions are:

- No electrodes and wires are attached to the driver.
- No cameras, monitors, light sources, or other devices are aimed at the driver (privacy issue).
- Less computational power is required for processing signals such as steering angle, which makes the online processing of data easily achievable.

- Hardware requirement for capturing signal from vehicle components such as steering, throttle, and gas paddle are much less than that required for an image processing or human body signals. These are often much cheaper and readily available.
- These methods are better suited for implementation due to their non-obtrusive nature.
- Due to variation in the dynamics of different types of vehicles, a universal system that will fit all vehicles is very difficult to achieve. These systems must be tuned in for the type of vehicle in use.
- Accuracy may not be high as compared to EEG monitors, since EEGs are constantly attached to the body and deliver a continuous signal (even when car is not in motion).

11 Design and Implementation Challenges to Consider

Generally, challenges associated with designing and implementing a drowsiness detection system are:

1. *Robustness.* The detection system should operate robustly in any driving condition. The design of the detection system has to be independent of environmental and non-drowsy-related factors that deviate vehicle performance state variable values from a normal driving baseline. For instance, one of the major obstacles in using steering wheel data for drowsiness detection is the dependency of steering values on road geometry and curvature. Some systems implement techniques that require eliminating the effect of curvature from the steering data (preprocessing) to handle the data independently of the road geometry. These methods have to be carefully assessed to avoid system/device false alarms and missed calls due to implementation of the preprocessor.

Naturally, each driver has his/her own individual style of driving and vehicle control. Some drivers are sensitive to lane position variations and make more small amplitude steering corrections to keep the vehicle in lane. Others are careless to their lane keepings and make less steering corrections with larger amplitude resulting in larger variations in vehicle lateral position. The performance of the drowsiness detection system may differ from driver to driver due to the variability in steering/vehicle control preferences among drivers. In addition, due to differences in vehicle dynamics and steering feel between passenger cars/trucks, steering ranges and variability are different as well. Therefore, a detection system using a vehicle-control-related variable has to be robust with respect to the aforementioned factors.

Lastly, environmental factors could highly affect the steering precision (Mackie and Wylie 1991). The designers should consider the effect of environmental and other external factors, that is, rain, fog, side-wind, speed, etc., on the performance of their proposed drowsiness detection system.

2. *Evaluation and User Acceptance.* Like other safety applications, drowsiness detection device need to go through extensive testing and validation procedures before full deployment. The detection system performance has to achieve a level of performance that can predict drowsy-related hazardous events accurately and timely with the minimum rate of missed and false alarms. Because of the risks of involving the drivers in dangerous drowsy scenarios, most evaluation results are based on prototype tests conducted in simulated or controlled environments. Simulator test results solely cannot provide compelling evidence that approves and guarantees a reliable operation of the device as a final product. However, simulation experiments are great tools for researchers and developers to test their algorithms' performance and identify the flaws associated to the proposed systems.

Field Operational Tests (FOTs) are the only validation method that can assess drowsiness detection systems as well as user acceptance in a naturalistic driving setting, while drivers have prolonged exposure to the technology (Barr et al. 2009).

Successful deployment of a system ultimately depends on end-user acceptance. Various studies show that the level of acceptance hinges upon the degree to which a driver perceives the benefits derived from a system as being greater than the costs. In other words, the potential benefits should outweigh the costs. On the other hand, if the users perceive significant enhancement in safety and driving skill, they might be encouraged to involve in riskier-than-usual driving behavior or become over-reliant on the system. Complete evaluation of the system requires full consideration of the aforementioned issues (Barr et al. 2009).

To evaluate the user acceptance, the following criteria have to be considered: 1) ease of use, 2) ease of learning, 3) perceived value, 4) driver behavior, and 5) advocacy.

❶ *Table 36.1* depicts a framework for user acceptance elements including the criteria for evaluation (Barr et al. 2009).

3. *Distraction.* The goal of driver assistant systems is to increase safety and decrease driver mental load which causes distraction and discomfort. Relatively, nonintrusive methods can get higher satisfaction rate among drivers with respect to intrusive methods using physiological parameters.
4. *Stakeholders buy-in.* All stakeholders (i.e., Departments of Transportation, trucking association, etc.) buy-in and acceptance have to be considered.
5. *Sampling rate.* Depends on the type of sensors used and detection algorithm developed. In majority cases the data is available at the desired sampling rate and is never considered a serious issue.
6. *Automatic vs. manual activation and deactivation.* This mode should be user selectable. Driver should have the choice to select automatic mode or the system is activated when he/she starts driving.
7. *Provision for privacy.* Designers and developers have to be aware of privacy concerns associated with systems that use re-identifiable personal information (i.e., face recognition methods).

Table 36.1
User acceptance elements as a part of a framework (Barr et al. 2009, Table 1)

Ease of use	Ease of learning	Perceived value	Driver behavior	Advocacy
Design and usability of device controls	Utility of instructions	Alertness and alertness management enhancement	Control inputs	Willingness to endorse
Use patterns	Ability to retain knowledge of device use	Driving skill enhancement	Awareness/behavioral adaptation	Purchase interest/intent
Understanding of device state	Time to learn	Safety	Driving style	
Driver accommodation/ variations in information processing		Health concerns	Lifestyle	
Demands on driver/ channel capacity		Confidentiality concerns		
Understanding of warnings/ discriminability of alerts				
Tolerance of false or nuisance warnings				

12 Conclusion

Drowsy and fatigue driving is a major transportation safety concern and is responsible for thousands of accidents and numerous fatalities every year. Driver fatigue is particularly significant in commercial vehicle operations because truck drivers stay on the road for extended period of time, which often involves driving at night. According to a recent CBS report drowsy driving costs the North American consumer \$16.4 billion in terms of property damages, health claims, lost time and productivity. Another \$60.4 billion/year are spent by US Government and businesses on accidents related to drowsy driving. The problem size is actually larger than most reported data because in most cases the cause is listed as something other than drowsiness. The problem significance and size is discussed in detail along with relevant statistics and references.

To mitigate the problem of driver drowsiness and reduce the number of accidents researchers have made tremendous efforts to develop systems that can help prevent such accidents. The most important challenge is to detect the driver’s condition sufficiently early, prior to the onset of sleep, to avoid collisions. Fatigue or drowsiness causes certain changes in the physiological and physical conditions of a driver. These changes also affect the performance and behavior of driver and are thus reflected in the output of vehicle he/she is driving.

Thus drowsiness or fatigue can be detected by recording one or more of these changes either directly from driver's body or from output of the driven vehicle. From the above discussion it is clear that detection systems can be broadly divided into two categories: (1) based on changes in physical and physiological changes and (2) changes in vehicle output.

Researchers have identified many variables in both categories. In the first category these include: changes in electric signals measured from human scalp also called Electroencephalogram (EEG); changes in eye closure activity measured through camera images or muscular activity around the eye; and changes in other variables such as heart beat, blood pressure, body posture, etc. These techniques are discussed in detail.

In the second category are changes in variables that are measured from output of the vehicle. These changes are a direct consequence of the drivers' inputs through vehicle controls. These controls include steering wheel, accelerator, and brake; the corresponding vehicle outputs are vehicle lateral position in the traffic lane, speed, acceleration, yaw rate, etc. By measuring changes in these variables one is actually measuring changes in driver behavior or state of alertness. Detection techniques based on vehicle control variables are the focus of this chapter. These are discussed in great detail along with examples of various systems based on these techniques.

Finally, issues related to design and implementation of detection systems in actual vehicles are also discussed.

The systems discussed here are mostly in research and prototyping stages. There are some commercially available systems but most of them are not properly validated. Some of them are meant for specific environment, some for specific vehicles. Still there are no such systems that are independent of environment, driver, vehicle, and other conditions. There is still some time before robust systems, that can detect drowsiness unobtrusively, are available to general public.

References

- Akerstedt T, Kecklund G (1994) Work hours, sleepiness and accidents. Karolinska Institute, Stockholm:1994, 104 (Stress Research Report No 248)
- Andreeva E, Arabi P, Philastides M, Mohajer K, Emami M (2004) Driver drowsiness detection using multi-modal sensor fusion. *Proc SPIE* 5434:380–390
- Artaud P, Planque S, Lavergne C, Cara H, de Lepine P, Tarriere C, Gueguen B (1994) An on-board system for detecting lapses of alertness in car driving. In: *Proceedings of the 14th E.S.V. conference session 2 – intelligent vehicle highway system and human factors*, Munich
- Barr L, Popkin S, Howarth (2009) An evaluation of emergin driver fatigue detection measures and technologies. Federal Motor Carrier Safety Administration, Report No: FMCSA-RRR-09-005
- Bergasa L, Barea R, Lopez E, Escudero M, Boquete L, Pinedo J (2003) Facial features tracking applied to driver drowsiness detection. In: *Proceedings of the 21st IASTED international conference on applied informatics*, Austria, Feb 2003, pp 378–825
- Brown ID (1966) Effects of prolonged driving upon driving skills and performance of a subsidiary task. *Ind Med Surg* 35:760–765
- Campagne A, Pebayle T, Muzet A (2004) Correlation between driving errors and vigilance level: influence of the driver age. *Physiol Behav* 80: 515–524

- Carskadon M, Dement W (1981) Cumulative effects of sleep restrictions on daytime sleepiness. *Psychology* 18:107–118
- Dinges DF (1995) An overview of sleepiness and accidents. *J Sleep Res* 4(Suppl 2):4–14
- Dingus TA, Hardee L, Wierwille WW (1985) Development of Impaired Driver Detection Measures. Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University, (Departmental Report 8504), Blacksburg
- Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
- Elling M, Sherman P (1994) Evaluation of steering wheel measures for drowsy drivers. In: 27th ISATA, Aachen, Germany, pp 207–214
- Eskandarian A, Mortazavi A (2007) Evaluation of a Smart algorithm for commercial vehicle driver drowsiness detection. In: IEEE Intelligent Vehicle Symposium, Istanbul, Turkey
- Eskandarian A, Mortazavi A (2008) An unobtrusive driver drowsy technology, provisional patent, Serial No. 61/193191, applied 5 Nov 2008, currently patent pending
- Eskandarian A, Mortazavi A (2009) Unobtrusive driver drowsiness detection system and method, US Patent Application No.12/613,306, filed 5 Nov 2009
- FARS (2008) National Highway Traffic Safety Administration, National Center for Statistics and Analysis, FARS Analytic Reference Guide, 1975 to 2008. US Department of Transportation, Washington, DC
- Fell D (1994) Safety update: problem definition and countermeasure summary: fatigue. New South Wales Road Safety Bureau, RUS No. 5
- Flandrin P, Goncalves P (2004) Empirical mode decompositions as data-driven wavelet-like expansions. *Int J Wavelets, Multi-Resol Inform Process* 2(4):477–496
- Fukuda J, Akutsu E, Aoki K (1995) Estimation of driver's drowsiness level using interval of steering adjustment for lane keeping. *JSAE Rev, Soc Automot Eng Japan* 16(2):197–199
- Gander P, James I (1999) Investigating fatigue in truck crashes. Report in Wellington school of medicine and Commercial Vehicle Investigation in New Zealand
- Harris W et al (1972) A study of the relationships among fatigue, hours of service, and safety of operation of truck and bus drivers, BMCS-RD-71-2. US Department of Transportation, Washington, DC
- Hertz RP (1988) Tractor-trailer driver fatality: the role of non-connective rest in a sleeper berth. *Accid Anal Prev* 20(6):429–431
- Horne J, Reyner L (1995) Sleep related vehicle accidents. *Brit Med J* 310(6979):565–567
- Huang NE, Shen SS (2005) The Hilbert-Huang transform and Its applications (Interdisciplinary mathematical sciences). World Scientific Publishing Company, Singapore, pp 324
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1996) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: Proceedings of the Royal Society A, Mathematical, Physical and Engineering Sciences, vol 454, London, UK, pp 903–995
- Hulbert S (1972) Effects of driver fatigue. In: Forbes TW (ed) Human factors in highway traffic safety research. Wiley, New York
- Jovanis P, Kaneko T, Lin T (1991) Exploratory analysis of motor carrier accident risk and daily driving patterns. 70th annual meeting of Transportation Research Board, Transportation, Washington, DC
- Lal S, Craig A (2001) A critical review of psychophysiology of driver fatigue. *Biol Psychol* 55: 173–194
- Mackie RR, O'Hanlon JF (1977) A study of the combined effects of extended driving and heat stress on driver arousal and performance. In: Mankie RR (ed) Vigilance: theory, operational performance and physiological correlates. Plenum Press, New York
- Mackie R, Wylie CD (1991) Countermeasures to loss of alertness in motor vehicle drivers: a taxonomy and evolution. In: Proceedings of the human factors society 35th annual meeting, San Francisco, CA, pp 1149–1153
- Mast T, Jones H, Heimstra N (1966) Effects of fatigue on performance in a driving device. *Highw Res Rec*. Cited in Haworth, Vulcan, Triggs, and Fildes, Driver fatigue research: development of methodology. Accident Research Center, Monash University Australia, 1989
- McCartt T, Ribner S, Pack A, Hammer M (1996) The scope and nature of the drowsy driving problem

- in the New York State. *Accid Anal Prev* 28:511–517
- McCartt A, Rohrbaugh J, Hammer M, Fuller S (2000) Factors associated with falling asleep at the wheel among long-distance truck drivers. *Accident Anal Prev* 32:493–504
- Mittler M, Miller J, Lipsitz J, Walsh K, Wylie C (1997) The sleep of long-haul truck drivers. *N Engl J Med* 337(11):755–761
- Mortazavi (2005) http://www.aaafoundation.org/multimedia/press_releases/PoliceDDFS.cfm
- Mortazavi A, Eskandarian A, Sayed RA (2009) Effect of drowsiness on driving performance variables of commercial vehicle drivers. *Int J Automot Engin* 10(3):391–404
- NASS/GES (2008) National Highway Traffic Safety Administration, National Center for Statistics and Analysis. National Automotive Sampling System (NASS): general estimates system, analytical user's manual 1988–2008. US Department of Transportation, Washington, DC
- Ogawa K, Shimotani M (1997) Drowsiness detection system. *Mitsubishi Electric Adv* 78:13–16
- Pilutti T, Ulsoy G (1997) Identification of driver state for lane-keeping tasks: experimental results. In: *The American Control Conference*, Seattle, pp 1667–1671
- Planque S, Petit C, Chapeau D (1991) A system for identifying lapses of alertness when driving. Renault, France
- Riemersma JB, Sanders AF, Wildervack C, Gaillard AW (1977) Performance decrement during prolonged night driving. In: Makie RR (ed) *Vigilance: theory, operational performance and physiological correlates*. Plenum Press, New York
- Rilling G, Flandrin P and Gonçalves P (2003) On empirical mode decomposition and its algorithms. In: *IEEE-EURASIP workshop on nonlinear signal and image processing*, grado (I), Baltimore
- Ryder J, Malin S and Kinsley C (1981) The effects of fatigue and alcohol on highway safety. NHTSA Report No. DOT-HS-805-854, Washington, DC
- Safford R, Rockwell TH (1967) Performance decrement in twenty four hour driving. *Highw Res Rec*
- Sagberg F (1999) Road accidents caused by drivers falling asleep. *Accident Anal Prev* 31:639–649
- Samel A et al (1995) Jet lag and sleepiness in aircrew. *J Sleep Res* 4(2):30–36
- Sayed R, Eskandarian A (2001) Unobtrusive drowsiness detection by neural network learning of driver steering. *J Autom Engin Proc Inst Mech Engin, Part D* 215:969–975
- Sayed (2011) <http://www.nhtsa.gov/FARS>
- Seki M, Shimotani M, Nishida M (1998) Study of blink detection using bright pupils. *JSAE Review, Society of Automotive Engineers of Japan* 19(1):58–60
- Seko Y (1984) Present technological status of detecting drowsy driving patterns. *Jidosha Gijutsu* 30(5), Central research Institute, Nissan Motor Company, Tokyo
- Siegmund GP, King DJ, Mumford DK (1996) Correlation of steering behavior with heavy-truck driver fatigue. *SAE Special Publications*, vol 1190, pp 17–38
- Skipper JH, Wierwille W and Hardee L (1984) An investigation of low level stimulus induced measures of driver drowsiness. Virginia Polytechnic Institute and State University IEOR Department Report #8402, Blacksburg, Virginia
- Stein AC (1995) Detecting fatigued drivers with vehicle simulators. In: Hartley L (ed) *Driver impairment, driver fatigue and driving simulation*. Taylor & Francis, Bristol, pp 133–150
- Stoohs R, Guilleminault C, Dement W (1993) Sleep apnea and hypertension in commercial truck drivers. *Sleep* 16:S11–S14
- Strohl KP et al (1998) NCSDR/NHTSA expert panel on driver fatigue and Sleepiness. drowsy driving and automobile crashes. Report and recommendations. National Highway Traffic Safety Administration, Washington, DC (DOT HS 808 707)
- Stutts J, Wilkins J, Osberg S, Vaughn B (2003) Driver risk factors for sleep-related crashes. *Accident Anal Prev* 35:321–331
- Sweeney M, Ellingstad V, Mayer D, Eastwood M, Weinstein E, Loeb B (1995) The need for sleep: discriminating between fatigue-related and non-fatigue related truck accidents. In: *Human factors and ergonomics society meeting*, vol 2, Baltimore, pp 1122–1126
- Thiffault P, Bergeron J (2003a) Monotony of road environment and driver fatigue: a simulator study. *Accident Anal Prev* 35:381–391
- Thiffault P, Bergeron J (2003b) Fatigue and individual differences in monotonous simulated driving. *Personal Individ Diff* 34:159–176
- Tilley DH, Erwin CW, Gianturco DT (1973) Drowsiness and driving: preliminary report

- of a population survey. SAE, international automotive engineering congress, Detroit Michigan, 8–12 Jan 1973
- Torsvall L, Åkerstedt T (1988) Extreme sleepiness: quantification of EOG and spectral EEG parameters. *Int J Neurosci* 38(3–4):435–441
- Wierwille W, Ellsworth L (1994) Evaluation of driver drowsiness by trained raters. *Accident Anal Prev* 26(5):571–581
- Wierwille WW, Wreggit SS, Mitchell MW (1992) Research on vehicle based driver status/performance monitoring. First Semi-Annual Research Report. NHTSA Cooperative Agreement Number DTNH 22-91-Y-07266, Washington, DC
- Wu R, Lin C, Liang S, Huang T, Chen Y, Jung T (2004) Estimating driving performance based on EEG spectrum and fuzzy neural network. *IEEE Int Conf Neural Netw* 1:585–590
- Yabuta K, Iizuka H, Yanagishima T, Kataoka Y, Seno T (1985) The development of drowsiness warning devices. In: *Proceedings of the 10th international technical conference on experimental safety vehicles*, Washington
- Young T, Blustein J, Finn L, Palta M (1997) Sleep-disordered breathing and motor vehicle accidents in a population-based sample of employed adults. *Sleep* 20(8):608–613

37 Drowsy and Fatigued Driver Warning, Counter Measures, and Assistance

Riaz Akbar Sayed¹ · Azim Eskandarian² · Ali Mortazavi³

¹Mechanical Department, NWFP University of Engineering and Technology, Peshawar, North West Frontier Provi, Pakistan

²Center for Intelligent Systems Research, The George Washington University, Washington, DC, USA

³Partners for Advanced Transportation Technology (PATH), University of California, Berkeley, CA, USA

1	<i>Introduction</i>	977
2	<i>Alarm Modality</i>	979
2.1	Visual Display	979
2.1.1	Warning Symbols	979
2.2	Auditory (Sound)	980
2.2.1	Warning Compliance (Urgency)	980
2.2.2	Tones and Voice Messages	981
2.2.3	Limitations of Auditory Warnings	981
2.3	Haptic/Tactile	982
3	<i>Alarm Timing</i>	983
4	<i>System Reliability and Sensitivity</i>	984
4.1	False Alarms	984
4.2	Setting Decision Thresholds for Warning Systems	984
4.3	Alarm Based on Likelihood of Falling Asleep	986
4.4	Graded (Staged) Alarm	987
5	<i>Designing a Warning System</i>	988
6	<i>User Acceptance and Trust</i>	990

7 **Counter Measures Against Driver Fatigue/Drowsiness990**

7.1 Legislation/Enforcement 990

7.2 Driver Education 991

7.3 Rumble Strips 991

7.4 Other Strategies 992

8 **Commercially Available Systems 993**

9 **Conclusion 993**

Abstract: Driving under the influence of fatigue and sleepiness is a serious safety concern. Hundreds of lives and billions of dollars are lost every year due to accidents caused by driver drowsiness. There are many aspects to the problem of a driver falling asleep while driving that include causes, detection, monitoring, warning, and countermeasures against drowsy driving. A number of crucial design issues have to be considered before the anticipated benefits of the drowsy driving warning can be fully realized. In this chapter the two major aspects, that is, warning and countermeasures, are discussed.

Warning means to convey to the driver about his/her state of sleepiness/drowsiness so that corrective actions can be taken. There are many issues related to warning system design but the two main concerns are when and how to warn the driver, that is, alarm modality and alarm timing. Although there are no standard guidelines for selection and design of appropriate alarm modalities, at least three types of modalities (visual, audio, and haptic/tactile) and their combinations are possible for any alarm design. An important component of collision avoidance system is the algorithm that determines the timing of warning. A poorly timed alarm may actually undermine the safety of the driver. An alert issued too early may be ignored by drivers if they are unable to perceive the cause of the warning. On the other hand, if it occurs too late, it may be viewed as ineffective.

An alarm that does not represent the true state of driver drowsiness, that is, the driver is not drowsy but the system issues a warning, is called false alarm. False and nuisance alarms are a particular problem for automotive collision avoidance and warning systems.

A comprehensive review of the literature regarding driver fatigue/drowsiness warning research, the present state of research and technologies being developed, and issues related to warning/alarm design and the future trends are highlighted. Driver fatigue-related countermeasures are also discussed along with their merits and demerits.

1 Introduction

In the last 10 years (2000–2010), more than 12,500 people have lost their lives in crashes related to driver fatigue/drowsiness (Fatality Analysis Reporting System (FARS) of the US department of transportation). Driver fatigue is particularly significant in commercial vehicle operations because truck drivers stay on the road for extended periods of time, which often involves driving at night. According to some estimates, drowsy driving costs the North American consumer \$16.4 billion in terms of property damages, health claims, lost time, and productivity. Another \$60.4 billion/year are spent by the US government and businesses on accidents related to drowsy driving. The problem size is very large and is discussed in detail in another chapter.

To mitigate the problem of driver drowsiness and reduce the number of accidents, researchers have put tremendous efforts to develop systems that can help prevent such accidents. Such systems will consist mainly of three parts:

- A detection system (an algorithm that analyzes data from sensors and detects any onset of sleep)

- A warning or alarm that alerts the driver and conveys the information to the driver through appropriate medium
- Other non-technological countermeasures that help prevent the onset of sleep

For a drowsy driver warning system to indicate potential state of drowsiness reliably and consistently, an appropriate detection algorithm is a must. Many algorithms have been proposed based on changes occurring in driver physical and physiological conditions such as electric signals from brain activity (EEG), eye closures, heart rate, skin electrostatic potential, blood pressure, and others. Another set of detection algorithms are based on signals from vehicle control functions that include vehicle steering activity, vehicle speed, vehicle lateral position, yaw rate, acceleration, and braking activity. These detection systems are discussed in detail in another chapter. In this chapter issues related to the warning and alarming part of the system are discussed.

Pritchett (2001) has defined an alerting system in the context of aircraft cockpits as an “attention-director.” She claims that alerting alarms should act as a trigger to the pilot to start the diagnosis and, if necessary, the resolution processes. To effectively incorporate any warning systems in vehicles, the most effective configurations for triggering drivers to take actions must be determined. Relevant information should be brought to the attention of individual drivers to help in critical decision making.

Adding a collision warning system to vehicles is actually increasing the information processing load placed on drivers and if not properly designed could divert attention away from a hazardous situation at the wrong time and could actually increase the crash risk instead of reducing it.

A well-designed warning system may help in reducing accidents related to driver drowsiness if the driver could use the warning information in a timely and effective manner. A number of crucial design issues have to be considered before the anticipated benefits of the drowsy driving warning can be fully realized. There is a need for further research to address issues of timing, modality, false alarms, and potential operator reaction to in-vehicle warning or alerting systems. For example, if the overhead involved in processing the information is too high then the driver may, correctly, ignore the warning.

There are a variety of issues pertaining to designing of alarm/warning systems:

- Alarm modality: visual, auditory, haptic/tactile.
- Alarm timing.
- User acceptance and trust:
 - False alarms: alarms without corresponding system problems
 - Nuisance alarms: alarms indicating a potential problem in an unrelated context
 - Inappropriate alarms: cascading alarms indicating minor problems
- System reliability and sensitivity: an alerting system can lack reliability if it fails to indicate a legitimate hazard, or if it activates without a true need.

In this chapter, various issues related to the warning/alarm part of the drowsiness assistance system will be discussed in detail. A comprehensive review of the literature regarding driver fatigue/drowsiness warning research, the present state of research and

technologies being developed, and issues related to warning/alarm design and the future trends will be highlighted. Driver fatigue-related countermeasures are also being discussed along with their merits and demerits.

2 Alarm Modality

Driver workload may be impacted by alarm modality. Any warning modality selected must be such that it fulfills the requirements of all drivers irrespective of age, personality, gender, etc.; it should also integrate with other alarms in the vehicle. Although there are no standard guidelines for selection and design of appropriate alarm modalities at least three types of modalities (visual, audio, and haptic/tactile) and their combinations are possible for any alarm design.

2.1 Visual Display

Displays serve to communicate information and information is needed for decision making and performing the proper control action in a timely manner. In aircrafts, pilots prefer visual over auditory warnings when there is enough time to react.

In drowsy conditions, visual warnings displayed in the dash board are less effective because the driver is less attentive and focused due to the onset of sleep. These displays cannot wake a sleepy driver but may warn the driver of the likelihood of falling asleep. For example, if there is high likelihood of falling asleep then it can be communicated through a message or symbol in the dash board.

The warning is more effective if provided in a head up display (HUD) located close to the forward view. Integration of HUD image with forward view of the real world and other warning systems make it very desirable. Eye accommodation is much less that can particularly benefit older drivers.

HUD image also has certain problems that include requirement of eye fixation, degradation of visual display, etc. The in-vehicle HUD provides information to a driver who is attending to the forward view. It cannot alert or warn the inattentive, distracted, or drowsy driver of potential hazard.

2.1.1 Warning Symbols

The effectiveness of a warning depends upon the characteristics of the warning itself, of the recipient of the warning, and the situation involved. Warning signs are intended to communicate knowledge about potential hazards and how to avoid them but can also be used as reminders that a hazard is present. Factors such as font, message, and inclusion

of a symbol will affect how effective a warning is. Use of symbols is preferred by many warning systems designers for a variety of reasons:

- Symbols can quickly grasp attention and communicate a message.
- Reading disability or language barriers are easily avoided.
- Symbols are easy to remember.

Poorly designed symbols are not properly understood. Training individuals to understand the meanings of various warning symbols can greatly improve the comprehension of warning symbols (Lesch 2003). To improve the effectiveness of warning symbols it is recommended that proper training should be provided to users. Sometimes pairing another piece of information such as associated text regarding the hazard also improves the comprehension of symbols.

2.2 Auditory (Sound)

Much of the information used by drivers is presented visually; therefore, it may be preferable to use alarms that access a different modality. Audio alerting signals might be best suited for automobiles since an automobile driver needs virtually constant eye contact with the road in order to maintain proper lane position. Replacing traditional visual indicators with auditory signals, such as bells, beepers, and electronic tones may reduce the need for visual instrument scanning, thereby allowing the user to devote more attention to other visual tasks. Auditory displays also have the advantage that they do not require the user, once alerted, to adjust his or her gaze in order to receive the message. Thus, they would be valuable in situations such as drowsy driving, where the driver is not able to focus attention due to onset of sleep. Landstorm et al. (1999) investigated the waking effect of sound in the driving environment and identified six important conditions that need attention:

1. High-frequency condition
2. Audibility condition
3. Tonal exposure condition
4. Acoustic pressure level condition
5. Variability condition
6. Disharmony condition

2.2.1 Warning Compliance (Urgency)

Warning compliance refers to the driver's understanding of the true message and level of urgency carried by the warning signal. Lack of warning compliance is also a big concern.

According to the Urgency Mapping Principle (Edworthy 1994) the urgency of the situation should match the perceived urgency of the alarm. Urgency of sound may guide people to give high priority to high urgent events. In driving environment appropriate

urgency mapping is necessary to enhance warning interpretations. High-urgency situations such as collision avoidance or drowsy warning should be mapped to a high-urgency sound and a low-urgency situation such as seat belt should be mapped to a low-urgency sound. An inappropriate mapping can undermine the importance of the warning and result in inappropriate response and may contribute toward nonacceptance of alarm systems.

Researchers have proposed many ways to heighten perceived urgency of alarm signals by manipulating signal parameters. Warning urgency could be increased by changing the following parameters:

- Fundamental frequency (high-frequency sound is perceived as high urgency)
- Sound pressure level (raising sound pressure also increases urgency)
- Inter-pulse intervals, that is, break between two successive tones (reducing inter-pulse intervals also increases level of urgency)

Warning design should consider how sound parameters affect urgency, annoyance, and appropriateness. Alerts whose meanings are not clear could delay response and increase the mental work load. A trade-off should always be made between urgency and annoyance or acceptance.

2.2.2 Tones and Voice Messages

- Auditory alarms can provide directional cues and according to Webber et al. (1994) auditory stimuli are processed faster than visual stimuli.
- Under normal conditions tones require less attention than voice but voice on the other hand can convey more detail information.
- Tones are independent of language.
- According to Edman (1982) speech may be more effective in high stress situations.
- Auditory icons which match driver's mental model provide faster more appropriate responses.

2.2.3 Limitations of Auditory Warnings

Auditory displays do possess certain limitations that also need to be considered:

- While the use of auditory information may help to alleviate the visual clutter, auditory displays by their very nature can be intrusive and distracting (Stokes et al. 1990).
- Drivers may get startled, annoyed, or both by auditory warnings especially for nonemergency situations or excessive false alarms.
- Drivers with hearing impairment (mostly old age) are not properly accommodated.
- Tones are not able to convey detailed information while voice is language dependent.
- Under drowsy or high ambient noise condition, signal deletion may cause problem.
- Algorithm with other alarms could distract driver with large number of verbal tones and messages.

2.3 Haptic/Tactile

Haptic or tactile warning refers to signals sensed through touch or body contact, for example, vibration-imposed signal in the steering wheel, driver seat, etc. Compared to an auditory interface, a tactile interface may be similarly salient with less disruption, but also may be less informative since instructive messages cannot be given. Tactile interfaces may also be preferable to drivers, as the alarms are less obvious to others in the car, and therefore less embarrassing to the driver (Dingus et al. 1998). Haptic displays are a promising way to present the warning message to drivers. According to many researchers drivers prefer haptic warning as compared to auditory. Haptic warning is considered less annoying and more trusted.

They have many advantages and certain short comings:

- Less mental load on drivers.
- Possibly shorter driver reaction time.
- Passengers are not aware of the warning.
- Haptic warning sensed through touch or body contact does not require any particular orientation.
- Haptic cues provide the quickest feedback to the operator and produce the fastest reaction time (Lloyd et al. 1999).
- Haptic cues do not require a specific orientation of sensory receptors for detection and are perceived very quickly.
- Haptic cues are not affected by most disabilities and impairments and can be detected by the majority of population.
- Haptic displays are intrusive and cannot be shut out easily.
- Although effective in gaining driver attention, haptic cues can convey only limited information.
- When used in an imminent warning situation they should supplement the main display.

Haptic communication to driver should be conveyed through the medium that the driver is in touch with such as accelerator, steering wheel, and driving seat. Haptic warning must be associated with proper driver response and must match driver's mental model of the situation.

Suzuki and Jansson (2003) tested subjects with sound and haptic warnings and found that steering vibrations are very effective for warning in lane departure situations (most common indicator of sleep onset). Many drivers have their own mental model for response to a haptic stimulus transmitted through steering wheel. This mental model causes drivers to think that the vehicle is deviating from the lane and not performing normally when steering vibrations are used as the warning signal.

Allowing the driver to choose the modality is another option.

Still another option is a multimodality alarm system.

3 Alarm Timing

An important component of collision avoidance system is the algorithm that determines the timing of warning. A poorly timed alarm may actually undermine the safety of driver. An alert issued too early may be ignored by drivers if they are unable to perceive the cause of the warning. On the other hand if it occurs too late, it may be viewed as ineffective.

Generally the driver has a very short period of time to wake up and take corrective action, which is why alarm timing is very critical and according to Janssen and Nilson (1993) it is a crucial factor in determining alarm effectiveness. One can assume that if alarm is provided early, the driver will have more time to respond and take corrective action. Early alarms are more helpful and effective than late alarms. It can be argued that the effectiveness of the alarm is directly proportional to the timing, the earlier the better. But early alarm corresponds to setting the threshold at a lower level, which will result in high rate of false alarms. To avoid the undesirable rate of false alarms and design an optimal system, a compromise has to be made.

How well a system is accepted is highly dependent on false alarm rates even if it is not a true false alarm but is perceived as false. This perception is heavily dependent on the timing of alarm. If the alarm is too late and the driver has already recognized the danger before the alarm is presented or if the alarm is too early and the driver knows that he/she has to take corrective action but is too early then the alarm will be perceived as false. This perception of false alarm due to mismatch between the driver's understanding of the situation and alarm may have the same effect on the acceptance rating of the system as that of true false alarms (Wheeler et al. 1998). The relationship between alarm timing and perception is very important but still needs more research and understanding.

Many researchers suggested that user behavior toward an automated system largely depends on their trust in that system. Abe (2002) found that real false alarm leads to decreased trust and delayed response to alarms. Late alarms lead to decreased trust compared to early or middle alarms. The timing of the alarm influence, the driver's trust in the warning system, and driver trustworthiness decision about an alarm are based on the timing to a great extent irrespective of its validity and may negatively impact the system effectiveness. The driver may accept warning systems with middle alarm timing more readily (Abe and Richardson 2004).

Early warning is of greater benefit than a late warning but the operational implication of the benefits depends on how much the early warning increases the false alarm rate. Drivers respond to these warnings as an automation that redirects attention rather than automation that triggers a response. The warning affects driver response by redirecting driver attention to the road.

According to some, the benefits of early warning in providing drivers with additional time to interpret and respond to a situation may outweigh the costs associated with false alarm.

4 System Reliability and Sensitivity

Important design considerations with respect to warning systems is the degree to which they perform reliably and the degree to which drivers believe that an alarm from the system indicates an impending dangerous situation. An alerting system can lack reliability if it fails to indicate a legitimate hazard, or if it activates without a true need.

4.1 False Alarms

An alarm that does not represent the true state of driver drowsiness, that is, the driver is not drowsy but the system issues a warning, is called a false alarm. False and nuisance alarms are a particular problem for automotive collision avoidance and warning systems. An important and dangerous issue with false alarms is their potential for driver distraction causing drivers to redirect their attention from the primary task of driving.

Many proposed warning systems are vulnerable to false alarms. For example, in the aviation industry, the Traffic Awareness and Collision Avoidance System (TCAS) initially considerably suffered from high rates of false alarms that resulted in pilots mistrust and lack of usage (Wiener 1988). Based on the studies of pilots usage of TCAS, if a warning and alerting system is purely designed with excessive false alarms then it will result in disengagement of the system and users will find creative ways of disabling it (Satchell 1993). The drowsy driver and other ITS alarms should be faced with similar problems (Horowitz and Dingus 1992; Knipling et al. 1993).

Every detection scheme is based on the analysis of various parameters (predictors of drowsiness). Once a particular parameter reaches a certain threshold the algorithm issues a warning. Selecting the value of this threshold is particularly important in determining overall performance of the warning system. If the threshold is set too low (alarm goes off easily) then there will be too many alarms and most of them may be false. Similarly if the threshold is set too high then there will be too few alarms and the chances of missing some drowsiness signals will increase.

4.2 Setting Decision Thresholds for Warning Systems

Setting the threshold for generating a drowsy driving alarm limits the number of false alarms at the cost of missing a drowsy driving signal. An ideal alarm system will have very little or zero false alarm frequency and 100% alarm accuracy, that is, zero misses. A designer will strive to achieve these objectives of zero false alarms and zero misses but in reality a compromise has to be reached. Setting the threshold too high to avoid false alarms will make the system more acceptable to the driver but at the cost of missing detection signals. The cost of missing detection signal is obviously too high as it involves human lives but the frequency of occurrence of such a signal is extremely low. The probability of falling asleep while driving on the road is very low as such an event may occur only few times in the life of

an individual driver and the time period is spread over years and may be decades. At the same time setting the threshold too low will result in excessive false alarms rendering the system less effective as drivers will completely ignore it due to annoyance.

A good basis for selecting appropriate threshold that will balance both criteria, that is, low false alarms and early detection, can be found based on the signal detection theory (Swets and Pickett 1982).

Let S be the signal generated by detection algorithm representing drowsy driving. N is the signal generated by detection algorithm representing nondrowsy driving, and R represents the event when the system generates drowsy driving alarm. The probability that the system will generate a drowsy alarm given the driver is actually drowsy, that is, correct detection, is given by $P(R|S)$. Similarly the probability that the system will generate a drowsy alarm given the driver is not actually drowsy, that is, false alarm, is given by $P(R|N)$. These probabilities could be used to determine the accuracy of the alarm system by using standard formulation from signal detection theory (SDT) and the effectiveness of different alarm systems could then be evaluated (Farber and Paley 1993).

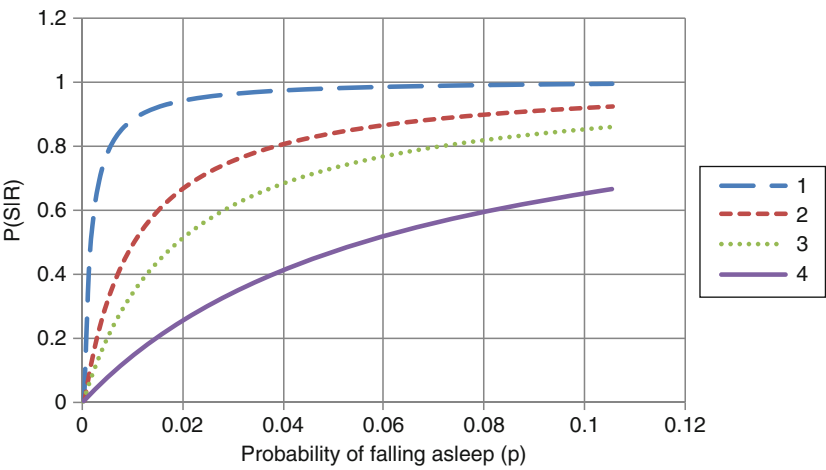
The probability that a driver will fall asleep while driving has a direct impact on the false alarm rate (Parasuraman et al. 1997). If p is the probability of falling asleep while driving at any given time, then the probability of falling asleep, given the system generates a positive alarm, is given by:

$$P(S|R) = \frac{P(R|S)}{P(R|S) + P(R|N) \frac{(1-p)}{p}} \quad (37.1)$$

Figure 37.1 is a graph of $P(S|R)$ and p for various values of accuracy and threshold values. Each curve represents a different level of threshold; these values are given in Table 37.1.

For the same accuracy the system false alarm rate varies considerably with the probability of falling asleep at the wheel. In Figure 37.1, each curve represents a different level of threshold setting. Curve number 1 represents a system in which the threshold is set very tightly with the probability of false alarms $P(R|N)$ of 0.001 and an accuracy $P(R|S)$ of 0.999. Even with such a strict threshold and high accuracy, a zero false alarm rate is almost impossible. Only when the probability of occurrence of drowsiness is in the range of 10% or above, the probability of true alarm $P(S|R)$ approaches one. Other curves are for more liberal thresholds, for example, curve 3 is for a system with $P(R|N)$ of 0.02 and $P(R|S)$ of 0.999. Although the probability of missing a drowsiness signal is much less but the probability of a true alarm $P(S|R)$ is much lower with high false alarm rate.

This could be clearly understood; for example, if the probability of the occurrence of sleep while driving (p) is 0.1% (0.001) for a system with detection accuracy $P(R|S)$ of 99.9% (0.999) and false alarm rate $P(R|N)$ of only 0.1% (0.001), then the chances of a true alarm are 44 out of 100. Similarly for the same system, if the probability of falling asleep at the wheel increases to 1% (.01) then the true alarm rate will increase to 88 out of 100. Because of the very low probability of falling asleep at the wheel, there will always be a false alarms no matter how accurate the system is.



■ Fig. 37.1
P(S|R) and p for various values of accuracy and thresholds

■ Table 37.1
Probability values for graphs in ● Fig. 37.1

Curve no.	Probability of false alarms P(R N)	System accuracy P(R S)
1	0.001	0.999
2	0.01	0.995
3	0.02	0.999
4	0.06	0.90

Setting the system threshold too high, resulting in very few false alarms, is not a very good idea (Farber and Paley 1993). False alarms are not particularly a bad thing. Since falling asleep while driving is such a rare event that if there are only true alarms a driver will never be familiar with it and will not know how to react and may react in a dangerous way. The probability of a true drowsy driving alarm may be low in the lifetime of a driver. False alarm rate within certain limits is a good thing and will make the driver familiar with the system. Of course too high a false alarm rate can annoy a driver and result in disbelief and low acceptance. Wheeler et al. (1998) suggest allowing drivers to adjust the sensitivity or threshold of the system to reflect the prevailing conditions, or the drivers’ own experience.

4.3 Alarm Based on Likelihood of Falling Asleep

Researchers argue that an ideal warning system will be one that generates alarm when there is likelihood of collision occurrence or onset of drowsiness knowing that the driver will avoid the danger by him/herself. Even though this is a false alarm, it will help the driver get familiar with the system and be able to respond in a much better way.

The probability (likelihood) of falling asleep while driving varies (McCartt et al. 2000). There are certain factors which increase the likelihood of falling asleep (likelihood of accident):

- Greater daytime sleepiness: more arduous schedules, with more hours of work and fewer hours off duty. The number and pattern of hours worked and hours off duty have been linked to sleepiness-related driving. Studies of the general driving population and shift workers have linked sleepiness-related driving with rotating shifts and night shifts (Mitler et al. 1988; Gold et al. 1992; Marcus and Loughlin 1996; McCartt et al. 1996; Lauber and Kayten 1998).
- Time of day: The time of day, has also been identified as predictive of sleepiness-related driving. Based on physiological and performance data of driver fatigue and alertness, the strongest and most consistent factor influencing driver fatigue and alertness is time of day (Mackie and Miller 1978; Wylie et al. 1996).
- Older, more experienced drivers.
- Shorter, poorer sleep on road: In general, a person's tendency to fall asleep during normal waking hours is increased and psychomotor performance declines with fewer hours of sleep and successive days of restricted sleep (Wilkinson et al. 1966; Carskadon and Dement 1981; Mitler et al. 1997).
- Symptoms of sleep disorder: Research suggests that drivers with un-treated sleep apnea, snoring, or sleep-disordered breathing are at increased risk for motor vehicle crashes (Findley et al. 1988; Aldrich 1989; Stoohs et al. 1993; Young et al. 1997).
- Greater tendency to nighttime drowsy driving (Hertz 1988; Jovanis et al. 1991).

All these and other factors (discussed in the chapter on detection) must be considered in determining the probability of the driver falling asleep at the wheel and then incorporating it in the warning algorithm. A properly designed system that has incorporated the above factors will thus have a very high true alarm rate at times when it is highly needed.

4.4 Graded (Staged) Alarm

In this case, the alarm signal is presented in a graded format based on the likelihood of the event. In case of rear-end collision, the proposed system is a graded sequence of warnings from mild to severe as a function of time to collision, "T." The longer T is, the milder the warning would be. If T is shorter than a critical time c, no warning would likely help the driver (Horowitz et al. 1992).

In the case of drowsy driving the graded sequence may consist of any of the three modalities changing from low urgency to high urgency as a function of severity of sleepiness (as determined by the detection algorithm).

Graded warning provides a greater safety margin and does not habituate drivers to warnings. Drivers trust graded warning more and the level of annoyance also does not increase (Lee and Hayes 2004).

There is some concern that staged (graded warning) may increase chances of false alarms and may not be very useful in highly dynamic and rapidly changing crash situation. The system may simply race through the stages directly to final stage of imminent crash. It may be more useful in slow speed environment in which the dynamics of the crash situation may allow drivers to benefit from earlier cautionary warnings.

5 Designing a Warning System

All kinds of alarms are not equally effective, many human factor considerations must be considered to design more effective alarms. Improperly designed warnings can add additional load to information processing and may divert driver attention. A number of crucially designed issues have to be considered before the anticipated benefits of the drowsy driving warning can be fully realized.

The warning design should consider how sound parameters affect urgency, annoyance, and appropriateness. Alerts whose meanings are not clear could delay response and increase the mental work load. A trade-off should always be made between urgency and annoyance or acceptance.

Horowitz and Dingus (1992) suggested the following approach.

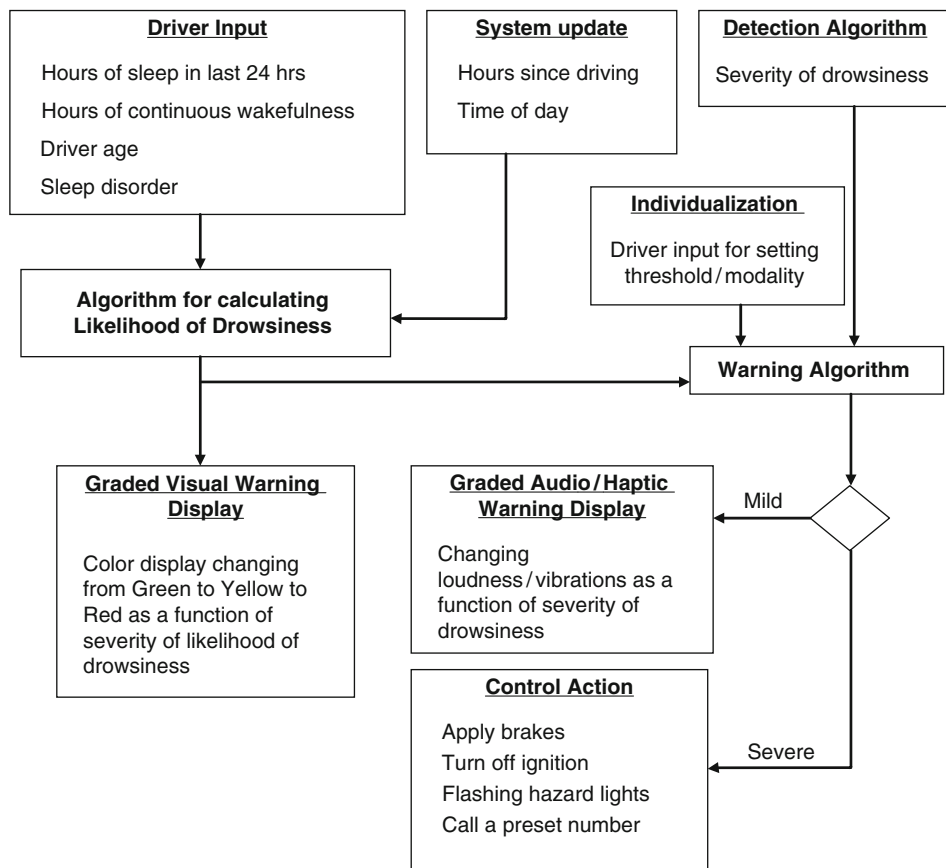
- Graded sequence of warning
- Parallel change in modality
- Individualization of warning

► *Figure 37.2* shows a schematic of a drowsy driver warning system based on the above approach.

At the start of the trip the driver enters the relevant information that can be predictors of drowsiness such as hours of sleep in last 24 h, hours of continuous wakefulness, driver age, and any sleep disorder. Hours since driving, and time of day are updated by the system. An algorithm then determines the likelihood of falling asleep from the above information. The output of the algorithm is conveyed to the driver through a graded visual color display. Based on the severity of the likelihood the color of the display will change from green to red. For example, if a driver starts his trip in the morning after a full night's sleep the display will show green but on the other hand if he starts at midnight with little sleep in last 24 h the display will be red, warning the driver that there is a high likelihood for him to fall asleep.

In parallel to the above, a drowsiness detection algorithm is also constantly analyzing the driver/vehicle condition to predict any onset of sleep. A warning algorithm, taking input from detection and likelihood algorithm, will activate an audio or haptic alarm in a graded sequence. If the driver does not respond and the detection algorithm outputs severe drowsiness along with high likelihood of drowsiness, then a control action by an automated system could also be considered that may include:

- Application of brakes
- Turn off ignition



■ Fig. 37.2
Schematic of a drowsy driver warning system

- Flashing hazard lights
- Call a preset number

The system can be individualized by setting the warning threshold of and warning modality by an individual driver based on his/her experience and tolerance with the system.

Overlapping two or more warnings can impair driver response. The algorithm should suppress the less urgent warning and allow only those alarms that are much more urgent, for example, a collision warning should proceed while blocking a telephone call. But this may also affect the acceptance ratings of the alarm system specifically when the false alarm rate is high as people do not like interception in their phone calls, etc.; so a compromise has to be reached.

All these and other factors (discussed in the chapter on detection) must be considered in determining the probability of driver falling asleep at the wheel and then incorporated in the warning algorithm. A properly designed system that has incorporated the above factors will thus have a very high true alarm rate at times when it is highly needed.

6 User Acceptance and Trust

Driver trust in automated systems is a major variable in determining the effectiveness of the system. Trust is highly related to use, misuse, and disuse of automated systems (Parasuraman and Riley 1997).

The extent to which the alerting system reliably indicates dangerous situations along with the degree of danger may impact the degree to which drivers trust in and utilize the system. Researchers have suggested that trust can affect how much people accept and rely on increasingly automated systems (Sheridan 1988; Lee and Moray 1992; Parasuraman et al. 1993; Muir and Moray 1996). Generally, research from both social science and engineering perspectives agree that trust is a multidimensional, dynamic concept capturing many different notions such as predictability, dependability, faith (Rempel et al. 1985), competence, responsibility, reliability (Muir and Moray 1996), robustness, familiarity, understandability, explication of intention, usefulness, and dependence (Sheridan 1988).

In a different environment, research has investigated driver trust and self-confidence in a simulated, in-vehicle decision aid which provided drivers with traffic information of varying degrees of reliability (Kantowitz et al. 1997). Drivers expressed less trust in the aiding system for conditions when the information was less reliable.

7 Counter Measures Against Driver Fatigue/Drowsiness

Countermeasures can be educational, cultural, and habitual in work and sleep habits, and legal in terms of regulating commercial vehicle operators, etc., that require education, awareness, and intervention at the societal level. They could also be technology driven that can be implemented as in-vehicle or highway (infrastructure) safety systems. These are discussed in the following sections.

7.1 Legislation/Enforcement

In 1938, the US congress enacted the federal hours-of-service (HOS) regulations applied to interstate commercial motor vehicles (CMV). These regulations limit the CMV drivers driving for 10 consecutive hours. Drivers are permitted to begin driving again for another 10 h after having 8 h off duty. In this way, the driver actually can drive a total of 16 h in a 24 h period. These rules were amended on January 1, 2004, and these new rules reduced the permissible hours of driving in each 24-h period from 16 to 14. These require 10 h of off duty time, which provides 2 h more for sleep and other hygiene functions. CMV drivers are required to complete their Record of Duty Status (RODS) also known as driver's log. However, according to some experts the HOS regulations are regularly flouted by a larger proportion of drivers particularly owner operators. One solution to this is the use of automatic recording of driver log using GPS, smart cards, and onboard monitoring.

In October 2002, the first bill regarding drowsy driving was introduced in the United States House of Representatives. The main feature of his bill is that it provides incentives to states and local communities for taking measures to enhance traffic safety related to driver fatigue. The bill calls for training of police officers, the creation of driver's education curriculum, standardized reporting of fatigue-related crashes on police report forms, and the promotion of countermeasures such as continuous shoulder rumble strips and rest areas.

In 2003, the state of New Jersey passed a legislation that established driving while fatigued as recklessness under the state's vehicular homicide statute. The legislation allows judges to consider drivers on trial for vehicular homicide as having driven recklessly, provided the driver had fallen asleep or had not slept for more than 24 consecutive hours before the accident occurred.

Radun and Radun (2009) reviewed the cases of fatigued or sleepy drivers who were convicted for drowsy driving offense in Finland. Despite the fact that it is difficult to prove the charges relating to drowsy driving, Finnish police and prosecutors were able to convict a significant number of drivers.

7.2 Driver Education

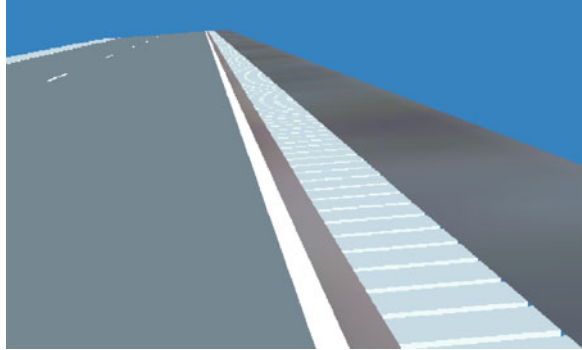
Educating drivers and creating awareness among the general population about drowsy driving is key to the solution of this problem. Introducing material regarding fatigue/drowsy driving and its countermeasures into driver training curriculum and driver licensing examination will be very effective in combating drowsy driving. The findings of Gander et al. (2005) suggest that fatigue management education is very useful for developing a fatigue management culture among CMV operators and that knowledge gained at the time of training is mostly retained and drivers implement the strategies against fatigue.

7.3 Rumble Strips

A rumble strip refers to a narrow band of unevenness in the road, normally placed on the shoulder (● Fig. 37.3).

When a wheel/tire rolls on this uneven surface it creates noise and vibration (rumble) in the vehicle. This rumble acts as an alarm and alerts the driver that he/she has departed the lane, which could be due to distraction or more commonly drowsiness or fatigue. The effectiveness of rumble strips has been studied by many researchers. Rumble strips at the center line can reduce the accidents by 15% and if also at the shoulder, reduction in accidents could be more than 40% (Presaud et al. 2003).

It is believed that when a sleepy driver hits a rumble strip the vibrations and noise produced has an alerting effect on drivers but it lasts for a short duration about less than 5 min and the drivers can go back to the same state of alertness as before hitting the strip. It could also be assumed that after hitting the rumble strip, the driver may react erratically due to sudden vibrations and noise but this normally does not happen, according to studies (Noyce and Elango 2004; Miles et al. 2006).



■ Fig. 37.3

Shoulder rumble strip (computer-rendered model)

Rumble strips may differ in types regarding length, width, depth, but there is no significant difference in the alerting effect of these different types (Anund et al. 2008).

7.4 Other Strategies

Strategies adopted by drivers in order to cope with fatigue and falling asleep at the wheel include a variety of activities that may stimulate the body and/or the mind. Most commonly used activities by professional drivers include (Royal 2003):

- Pulling over to take a nap (43%)
- Opening the window (26%)
- Drinking a hot or cold caffeinated drink 17%
- Getting off the road (15%)
- Turning on the radio (14%)
- Stretch or exercise (9%)
- Changing drivers (6%)
- Eating (3%)
- Singing or talking to self or others (3%)

Drinking an energy drink (high content of caffeine and taurine) prior to driving has a positive effect in counteracting fatigue in the short run but in the long run it has a negative rebound effect. Performing a mental dexterity task such as shelling and eating sunflower seeds during the drive seems to have the fatigue suppressing effect (Gershon et al. 2009).

These strategies or countermeasures can only delay, to a lesser extent, the onset of drowsiness but may not increase alertness when drowsiness has already been detected. According to an expert panel report (Strohl et al. 1998), these activities have no scientific basis to support their effectiveness except for taking a nap and caffeine intake that can provide a delayed and temporary relief.

8 Commercially Available Systems

For any drowsy detection system to be commercially successful, it has to be nondriver specific, nonintrusive, work in real time, be unobtrusive, and have no physical contact with the driver. It should also cause no harmful emissions or include any moving parts. The only commercially available device that is partially validated is the PERCLOS-based camera by Attention Technologies. Nissan, Ford, and Toyota all developed systems based on steering activity but are never implemented and still in research stage. Devices based on head dropping/nodding are also available from many vendors but are not validated.

9 Conclusion

Driver drowsiness is a serious safety issue that is responsible for thousands of traffic accidents and loss of precious lives and properties worth billions. There are many aspects to this problem including detection, monitoring, warning, and countermeasures to avoid fatigue/drowsiness. Warning and countermeasures are the two aspects that were discussed in this chapter. Warning means to convey to the driver about his/her state of sleepiness/drowsiness so that corrective actions can be taken. The main issues relating to warning are when and how to warn the driver, that is, alarm modality and alarm timing.


If the information is not conveyed through the proper medium, then the driver will either never get the information or never understand what the message conveyed to him/her is. For example, if a driver is falling asleep and the warning is conveyed through a visual display in the dash board then chances are that the warning will never make it to the driver. In the state of high fatigue/drowsiness the ability of individuals to focus attention and take control actions diminishes significantly and he/she may never scan the dash for warning display. In such case, a haptic or audio mode is highly desirable that has the ability to wake up the driver and convey the message that the driver is under the influence of fatigue/drowsiness. Visual display is good for alarming about the existence of the highly likelihood of occurrence of falling asleep at the wheel.

Another important aspect of warning a drowsy driver is the timing of alarm. If the alert is provided at the last time (too late) then there will be no time for the driver to take any corrective action. On the other hand, if the alert is provided way before the actual onset of sleep (too early) then there are chances that drivers will consider it as false alarm and not trust them. The issue of false alarms is very important and one of the major causes of warning system failures due to nonacceptance by drivers.

Setting threshold for issuing an alert is also an important aspect and if improperly set then many issues can arise. For example, if the threshold is set too low then there will be too many alarms and chances of false alarms will be high that lead to mistrust and unreliability. On the other hand, if the threshold is set too high then the chances of missing drowsiness signals will be high that can be very dangerous. The probability of falling asleep while driving is so low that it may not happen for years or decades; therefore, if there are no false alarms then the driver would never know what to do when it

sounds (warns) for the first time. Some false alarms are needed to keep the driver familiar with the system. Proper setting of threshold level is discussed in detail.

Designing a warning system for a drowsy driver is a very challenging task and requires knowledge of the above-mentioned parameters. A schematic layout for designing such a warning system is discussed in detail.

Not all systems are fit for all drivers; each driver has his/her own personality and choices. It is therefore important that some provision should be made for individualization of the warning system. In the schematic  Fig. 37.2 it is shown that allowing the individual driver to set certain parameters such as selecting modality, threshold value, timing, etc., can help to individualize the system.

Drowsy driver detection and warning is still in its infancy and research is continuing to develop more robust and reliable systems. The various aspects discussed here are mostly in research stages. Researchers are closing in on robust and reliable systems that can become available commercially and to the general public.

References

- Aldrich MS (1989) Automobile accidents in patients with sleep disorders. *Sleep* 12(6):487–494
- Abe G, Itoh M, Tanaka K (2002) Dynamics of driver's trust in warning systems. In: *Proceedings of the IFAC world congress, Barcelona*
- Abe G, Richardson J (2004) The effect of alarm timing on driver behavior: an investigation of differences in driver trust and response to alarms according to alarm timing. *Transport Res Part F* 7:307–322
- Anund A, Kecklund G, Vadeby A, Hjälm Dahl M, Åkerstedt T (2008) The alerting effect of hitting a rumble strip—a simulator study with sleepy drivers. *Accid Anal Prev* 40(2008): 1970–1976
- Carskadon MA, Dement WC (1981) Cumulative effects of sleep restriction on daytime sleepiness. *Psychophysiology* 18:107–118
- Dingus TA, Jahns SK, Horowitz AD, Knippling R (1998) Human factors design issues for crash avoidance systems. In: Barfield W, Dingus TA (eds) *Human factors in intelligent transportation systems*. pp 55–93
- Edman TR (1982) Human factors guide lines for the use of synthetic speech devices. In: *Proceedings of the human factor society 26th annual meeting, Human Factors Society, Santa Monica, 1982*. pp 212–216
- Edworthy J (1994) The design and implementation of nonverbal auditory warnings. *Appl Ergon* 25(4):202–210
- Farber E, Paley M (1993) Using freeway traffic data to estimate the effectiveness of rear end collision countermeasures. In: *3rd annual IVHS America meeting, IVHS America Washington DC*
- Findley LJ, Unverzagt ME, Suratt PM (1988) Automobile accidents involving patients with obstructive sleep apnea. *Am Rev Respir Dis* 138:337–340
- Gander P, Marshall N, Bolger W, Girling I (2005) An evaluation of driver training as a fatigue countermeasure. *Transp Res Part F Traffic Psychol Behav* 8:47–58
- Gershon P, Shinar D, Ronen A (2009) Evaluation of experience based fatigue countermeasures. *Accid Anal Prev* 41(2009):969–975
- Gold DR et al (1992) Rotating shift work, sleep, and accidents related to sleepiness in hospital nurses. *Am J Public Health* 82(7):1011–1014
- Hertz RP (1988) Tractor-trailer driver fatality: the role of non-connective rest in a sleeper berth. *Accid Anal Prev* 20(6):429–431
- Horowitz AD, Dingus TA (1992) Warning signal design: a key to human factors issue in an in-vehicle front to rear end collision warning system. In: *Proceedings of the human factors society, vol 36*, pp 1011–1013
- Janssen W, Nilson L (1993) Behavioral effects of driver support. In: Parkes AM, Fransen S (eds) *Driving future vehicle*. Taylor & Francis, London, pp 147–155

- Jovanis PP, Kaneko T, Lin TD (1991) Exploratory analysis of motor carrier accident risk and daily driving patterns. In: 70th annual meeting of Transportation Research Board, Transportation Research Board, Washington, DC, 1991
- Kantowitz BH, Hanowski RJ, Kantowitz SC (1997) Driver acceptance of unreliable trac information in familiar and unfamiliar settings. *Hum Factors* 39:164–176
- Knipling RR, Mironer M, Hendricks DL, Tijerna L, Everson JC, Wilson C (1993) Assessment of IVHS countermeasures for collision avoidance: rear-end crashes, NTIS No. DOT-HS-807-995. National Highway Traffic Safety Administration, Washington, DC
- Landstrom U, Englund K, Nordstrom B, Astrom A (1999) Sound exposure as a measure against driver drowsiness. *Ergonomics* 42(7):927–937
- Laubert JK, Kayten PJ (1998) Sleepiness, circadian dysrhythmia, and fatigue in transportation system accidents. *Sleep* 11:503–512
- Lee JD, Moray N (1992) Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35:1243–1270
- Lee JD, Hoffman JD, Hayes E (2004) Collision warning design to mitigate driver distraction, CHI 2004, Vienna, 24–29 Apr 2004
- Lesch MF (2003) Comprehension and memory for warning symbols: age-related differences and impact of training. *J Saf Res* 34:495–505
- Lloyd MM, Wilson GD, Nowak CJ, Bittner AC (1999) Brake pulsing as haptic warning for an intersection collision avoidance countermeasure. *Transp Res Rec* 1694:34–41
- Mackie RR, Miller JC (1978) Effects of hours of service, regularity of schedules, and cargo loading on truck and bus driver fatigue, DOT-HS-5-01142. Human Factors, Goleta
- Marcus CL, Loughlin GM (1996) Effect of sleep deprivation on driving safety in housestaff. *Sleep* 19(10):763–766
- McCartt AT, Ribner SA, Pack AI, Hammer MC (1996) The scope and nature of the drowsy driving problem in New York state. *Accid Anal Prev* 28(6):511–517
- McCartt AT, Rohrbaugh JW, Hammer MC, Fuller SZ (2000) Factors associated with falling asleep at the wheel among long-distance truck drivers. *Accid Anal Prev* 32(4):493–504
- Miles JD, Pratt MP, Carlson PJ (2006) Evaluation of erratic maneuvers associated with installation of rumble Strips. *J Transp Res Board* 1973:73–79
- Mitler MM, Carskadon MA, Czeisler CS, Dement WC, Dinges DE, Graeber RC (1988) Catastrophes, sleep, and public policy. *Sleep* 11(1):100–109
- Mitler MM, Miller JC, Lipsitz JJ, Walsh JK, Wylie CD (1997) The sleep of long-haul truck drivers. *N Engl J Med* 337(11):755–761
- Muir BM, Moray N (1996) Trust in automation: part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39:429–460
- Noyce DA, Elango VV (2004) Safety evaluation of centerline rumble strips. *J Transp Res Board* 1862:44–53
- Parasuraman R, Molloy R, Singh IL (1993) Performance consequences of automation-induced “complacency.” *Int J Aviation Psychol* 3:1–23
- Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 39(2):230–253
- Parasuraman R, Hancock PA, Olofinboba O (1997) Alarm effectiveness in driver centered collision warning systems. *Ergonomics* 40(3):390–399
- Presaud BN, Retting RA, Lyon CL (2003) Crash reduction following installation of centerline rumble strips on rural two-lane roads. Ryerson University, Toronto
- Pritchett AR (2001) Reviewing the role of cockpit alerting systems. *Hum Factors Aerosp Soc* 1:5–38
- Radun I, Radun JE (2009) Convicted of fatigued driving: who, why and how? *Accid Anal Prev* 41(2009):869–875
- Rempel JK, Holmes JG, Zanna MP (1985) Trust in close relationships. *J Pers Soc Psychol* 49:95–112
- Royal D (2003) National survey of distracted and drowsy driving attitudes and behavior: 2002 (vol 1: Findings), DOT HS 809 566. US Department of Transportation NHTSA, Washington, DC
- Satchell P (1993) Cockpit monitoring and alerting systems. Ashgate, Aldershot
- Sheridan TB (1988) Trustworthiness of command and control systems. In: Proceedings of the IFAC/IFIP/IFORS/IEA conference on analysis, design, and evaluation of man-machine systems, Oulu, Finland. pp 427–431
- Stokes A, Wickens C, Kite K (1990) Display technology: human factors concepts. Society of Automotive Engineers, Washington DC

- Stoohs RA, Guilleminault C, Dement WC (1993) Sleep apnea and hypertension in commercial truck drivers. *Sleep* 16:S11–S14
- Strohl KP, Blatt J, Council F (1998) Drowsy driving and automobile crashes. Report of NCSDR/NHTSA expert panel on driver fatigue and sleepiness, DOT HS 808 707. US Department of Transportation NHTSA, Washington, DC
- Suzuki K, Jansson H (2003) An analysis of driver steering behavior during auditory or haptic warnings for the design of lane departure warning system. *JSAE Rev* 24(2003):65–70
- Swets JA, Pickett RM (1982) Evaluation of diagnostic systems. Academic, New York
- Webber JW, Mullins CA, Schumacher PW, Wright CD (1994) A system approach to the development of an integrated collision avoidance vehicle. In: Proceedings of the vehicle navigation and information systems conference, pp 431–434
- Wheeler WA, Campbell JL, Kinghorn RA (1998) Commercial vehicle-specific aspects of intelligent transportation systems. In: Barfield W, Dingus TA (eds) Human factors in intelligent transportation systems. Erlbaum, Mahwah, pp 95–130
- Wiener EL (1988) Cockpit automation. In: Wiener EL, Nagel DC (eds) Human factors in aviation. Academic, San Diego, pp 433–461
- Wilkinson T, Edwards S, Haines E (1966) Performance following a night of reduced sleep. *Psychon Sci* 5:471–472
- www.nhtsa.gov/FARS
- Wylie CD, Schultz T, Miller JC, Mitler MM, Mackie RR (1996) Commercial motor vehicle driver fatigue and alertness study: technical summary, MC-97-001. Federal Highway Administration, Washington, DC
- Young T, Blustein J, Finn L, Palta M (1997) Sleep-disordered breathing and motor vehicle accidents in a population-based sample of employed adults. *Sleep* 20(8):608–613

Section 8

Vision-based Systems

Alberto Broggi

38 Image Processing for Vehicular Applications

Massimo Bertozzi

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Setup</i>	1000
1.1	Functionality	1000
1.2	Technical Feasibility of Device Positioning	1001
1.3	Wiring and Positioning	1002
1.4	Lighting Control	1002
2	<i>Specific Machine Vision Issues in the Automotive Field</i>	1002
2.1	Vehicle Ego-Motion	1004
2.2	Oscillations and Vibrations	1004
2.3	Illumination Conditions	1006
3	<i>Conclusions</i>	1008

Abstract: In developing a vision system for a vehicle, different setup constraints and issues must be considered.

Space, wiring, or lighting are also typical issues to be also faced in industrial scenarios; nevertheless, when a vision system has to be deployed inside a vehicle they have to be more carefully studied and often drive the hardware selection.

Moreover, cameras are to be installed on moving vehicles and this led to additional problems to be faced. In fact, camera movements, oscillations and vibrations, or different and even extreme illumination conditions have to be taken in account when developing machine vision software.

1 Setup

The development of machine vision systems requires to cope with typical problems like: background noise, camera movements, illumination conditions, characteristics of the target to detect, and many others. While these problems also affect industrial applications or video surveillance systems in the automotive field, they are largely amplified. In fact, the scenario changes continuously, the vision systems are installed on board of moving vehicles, the vehicle and therefore also the vision system are subject to vibrations and oscillations due to road coarseness, and some targets like pedestrians can only be defined in a statistical and non exhaustive way. Moreover, the sensor positioning is often constrained by wiring or design requirements and does not allow to select the optimal viewing position.

For these reasons, the setup design is one of the most complex challenges in the implementation of a complete system and system designers have to comply with the constraints discussed in the following section.

1.1 Functionality

Different ADAS are already available on the market and others will be available shortly (Norén 2008): Adaptive Cruise Control, All-Round-View, Collision Warning and Auto Brake, Precrash Safety, Lane Departure Warning, Lane Keeping Assistant, Stop-and-Go Assistant, Blind Spot Detection, Lane Change Assistant, and Night Vision.

The hardware setup strongly depends on the specific functionality. Some of these systems like Lane Departure Warning or Blind Spot Detection require a simple hardware setup: one smart camera connected to an integrated display on the vehicle. While other systems like Stop-And-Go or Collision Warning and Auto Break require a more complex setup: a stereoscopic system or a sensor fusion with other devices.

ADAS providing complex precrash features, such as pedestrian detectors, require a more complex design since they need to process data from several sensors which might already be used for other purposes –such as a single wheel speed detector for ESP (Electronic Stability Program)– to perform their task.

For multi-sensor systems, synchronization must be ensured to avoid artificial data realignment inside the ECU (Electronic Control Unit). Synchronization must be supported by sensors, and is usually distributed as a square wave triggering the sampling instant. If sensors only supply a strobe signal, a robust time-stamping inside the ECU is required in order to allow real-time data alignment. Data alignment can be problematic from some sensors like forward looking cameras.

1.2 Technical Feasibility of Device Positioning

In the prototyping phase, sensors installation must follow a feasibility analysis. During this phase, constraints like installation cost and system performance must be considered together with esthetics or ergonomics. The perception system components can be placed all around the vehicle depending on the application without limiting the visibility for the driver, and can be placed both inside or outside the cabin.

These choices are driven by both the target application and technological issues. Inside the cabin the camera is protected from rain, snow, and dust but has to follow some esthetic and ergonomic constraints. Moreover, modern heat-treated windshields filter the near infrared wavelength causing loss of information if the system uses an infrared camera sensor. This problem can be solved in different ways such as replacing the windscreen or moving the camera (i.e., thermal cameras) outside the cabin.

Far infrared cameras cannot be placed inside the cabin since glass is opaque to these wavelengths. ▶ [Figure 38.1.a](#) shows an example of FIR camera integration. However, an outdoor installation has to cope with environment-related problems such as cleaning the device, waterproof resistance, and in some cases shock resistance. Devices mounted in peripheral positions –such as behind the bumper– need a protection system from shocks. ▶ [Figure 38.1b](#) shows a possible solution for the camera setup in a Start-Inhibit system on a truck (Broggi et al. 2007).

**a****b**

■ Fig. 38.1

Example of the integration of a FIR vision system: an infrared camera is mounted in a central position in the front of the vehicle (a). Integration of a stereo vision system on a truck (b)

1.3 Wiring and Positioning

Wiring and positioning of both sensors and processing engines have to be carefully considered when needed.

In order to minimize the esthetic impact on commercial vehicles and to ease the vision systems integration, the use of small cameras is mandatory for ADAS systems. At the same time, the need of a reasonable processing power to perform recognition tasks must be fulfilled as well.

The use of standard industrial smart cameras that include sensor and processing unit in a single enclosure, while generally fulfilling the processing power requirement, does not fit the size constraints. A widely adopted solution is based on using two separated vision sensor and processing units. The units are connected by some robust interface such as: Ethernet, Firewire, or USB busses. In this way the embedded processing unit of appropriate power can be placed more freely on the vehicle where there is available space. Moreover, a single processing unit can be used for different ADAS processings. Some systems may have the ECU placed in proximity of the sensor and produce the results – such as driver warnings – directly there. However if sensors are placed outside the cabin, connection cables between the sensor and the ECU must be placed taking into account problems such as temperature range, electromagnetic interferences generated by the engine, and thermal noise, all of which cause signal degradation. These problems are critical if the signal has a high frequency, such as for high resolution or high frame rate cameras. Differential buses such as CAN (Controller Area Network), Firewire, or LVDS (Low Voltage Differential Signal) provide the necessary robustness for communication.

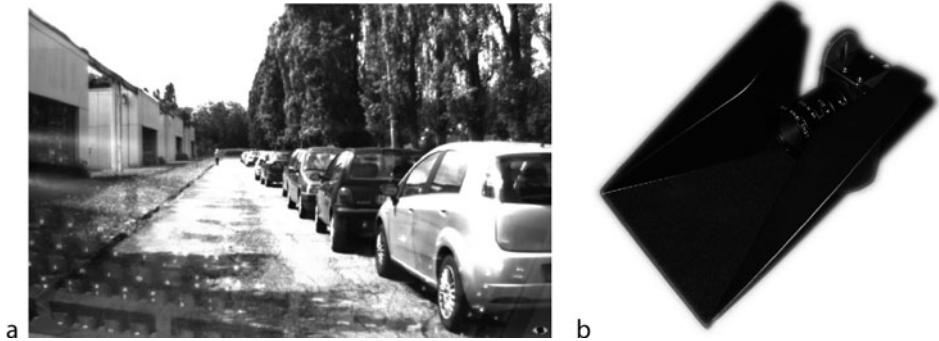
1.4 Lighting Control

During the day, scene illumination is determined by weather conditions. When the camera is placed inside the cabin, internal illumination can cause reflections on the glass (see ● [Fig. 38.2a](#)); to avoid this effect a small black chamber can be installed around the camera (like the one in ● [Fig. 38.2b](#))

At night, on the other hand, illumination is poor even with a NIR camera and the system requires a proper illuminating hardware (with respect to the camera sensitivity spectrum). In ● [Fig. 38.3.a](#), setup with two different NIR lamps is shown. In ● [Fig. 38.3b](#), the NIR illuminator has been integrated within the head lamp assembly.

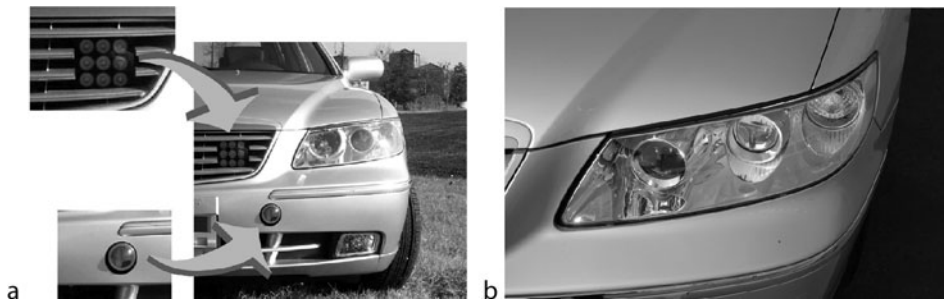
2 Specific Machine Vision Issues in the Automotive Field

As previously introduced, issues specific to the automotive field must be faced in developing vision-based ADAS systems. The main issue is due to the fact that cameras are



■ Fig. 38.2

Color image acquired by an onboard camera with reflections in the bottom (a). A possible solution to protect the camera sensor from reflections (b)



■ Fig. 38.3

Two different types of near infrared lamps installed on an experimental vehicle (a). Night vision lighting system composed by low/high beam light to the *left* and NIR lamp in the *middle* or (b) integrated within the head lamp

installed on moving vehicles and therefore the vision system and its related processing steps must be robust with respect to vehicle movements. In most cases, the vehicle's ego-motion must be taken into account. Besides ego-motion, also other kinds of movements like vibrations or oscillations represent a source of noise for vision-based systems.

Other issues are related to specific environmental conditions in outdoor environments. In fact, temperature and illumination conditions can vary and can be barely controlled. Especially for illumination, extreme situations like direct sunlight or strong reflections must be taken into account. In addition, other light sources, such as car headlights or reflectors, might be present in a typical automotive scene.

Specific camera issues related to the automotive environment are summarized in [Table 38.1](#).

■ Table 38.1
Common automotive applications with typical camera features

Issue	Ego-motion	Oscillations and vibrations	Illumination conditions
Properties	Moving background, perspective changes	Noise overlapped to ego-motion	Object texture changes, bad reflections
Impact	Motion blur, object changes	Tracking problems	Camera dazzling, bad recognition
Workaround	Faster shutters, better processing	Better ego-motion detection	Better processing, higher dynamic range

2.1 Vehicle Ego-Motion

When the vision system is installed on board of a vehicle, it has to be robust with respect to vehicle movements. This design issue can be examined at two different levels: vision devices (i.e., cameras configuration) and processing (algorithms).

Concerning the cameras, some technologies are not robust to motion artifacts, that is, moving objects are blurred in acquired images. This effect is particularly evident when the vehicle makes a sharp turn, and the whole background begins to move. ➤ [Figure 38.4](#) shows the effect for a FIR camera.

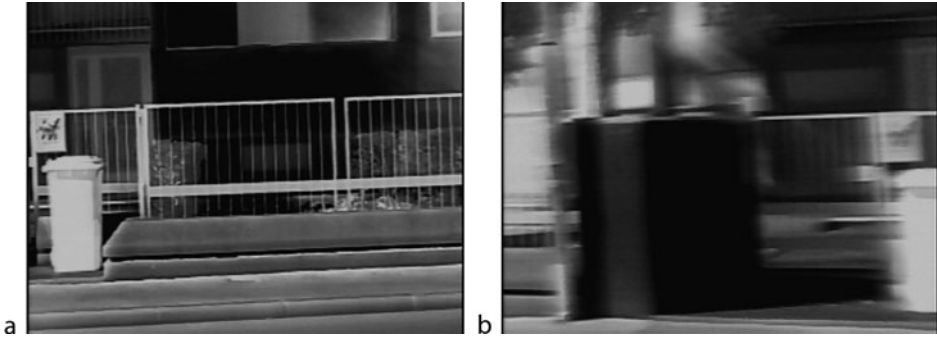
While blurring can even help in some scenarios and for specific vehicle movements by hiding unnecessary details, generally it has to be avoided. Therefore, a careful camera selection is a mandatory step in designing the setup. Specifically, old CMOS-based cameras more likely feature a slow sensor and, thus, can be affected by this problem. Conversely, the effect is not appreciable for CCD-based cameras and, generally, in recent CMOS models.

Vehicle movements, namely ego-motion, must be considered as input for many image processing algorithms. The computation of ego-motion for a vision system can be performed using machine vision techniques like the analysis of background movements or visual odometry (Dickmans and Mysliwetz 1992); however, these techniques require additional computing and are not always applicable, in such cases added (and often expensive) sensors like gyroscopes, odometers, or inertial devices are generally used.

2.2 Oscillations and Vibrations

Oscillation and vibrations have been already discussed from a mechanical point of view for calibration, in this section, specific issues for processing in automotive applications are covered.

Besides tracking, other vision-based applications are affected by vehicle movements as well. In fact, many systems rely on calibration to recover 3D information or to detect objects. Unfortunately, vibrations or oscillations, induced by normal vehicle functioning, affect calibration and may lead to incorrect results.



■ Fig. 38.4

An example of motion blur in FIR images acquired by an onboard vision system: the left image was captured when the vehicle was still, while the right snapshot was taken only a few seconds later –when the vehicle turned left– and shows a heavy horizontal motion blur effect

Therefore, image stabilization techniques are widely used to cope with this problem. In some cases, this can be done during the acquisition step, since some cameras feature image stabilization at sensor level. Another hardware-based solution is the use of electromechanical stabilization platforms (Schiehlen and Dickmanns 1994) or lenses-based mechanisms (Cardani 2006). These approaches are generally effective for suppressing really abrupt movements but are less suited for removing the specific range of movements due to oscillations or vibrations typical of the automotive field (Bombini et al. 2006).

In most situations, a specific processing phase devoted to removing this source of noise has to be developed. This is a difficult task, since only unwanted motions have to be removed, while the motion components due to vehicle ego-motion have to be preserved.

Vibrations and oscillations are considered the *high-frequency* component of global motion and therefore image stabilization can be applied in an attempt to smooth inter-frame motions. In specific situations, this task can be simplified in order to remove critical noise components only; in fact, the definition of *unwanted motions* can depend on the specific application; as an example, pitch variations can highly affect distance estimation for monocular systems which often relies on the vertical features positioning in the acquired images to estimate distance; in such a specific case only pitch deviations should be removed to avoid wrong distance estimation (Bombini et al. 2006). Conversely, in a stereo system, distance can be computed exploiting 3D triangulations but, at the same time, a large number of stereo vision-based systems are based on the assumption of a null roll. In such cases, pitch oscillations barely affect the processing while roll variations have to be compensated.

An image stabilization process is generally divided into two different steps: inter-frame motion detection and motion compensation.

In the first step, most systems exploit feature detection and tracking to recover motion. Again, the nature of the features to extract highly depends on the stabilization requirements: for simple stabilization techniques or when real-time constraints apply, simple features are generally extracted like edges (Bombini et al. 2006). More complex features, like lane markings, are used when a more precise stabilization process is required (Liang et al. 2003).

A different approach for motion detection is based on the use of dense matching techniques like image disparity or optical flow computation.

The motion compensation stage is used to compute the roto-translation which is applied to consecutive frames to minimize noise introduced by vibrations and oscillations. In simple cases, it is based on a low-pass filter to remove high frequency components of movements, but also more complex approaches that exploit supplementary information on the scenario, like object or background position, are widely used.

2.3 Illumination Conditions

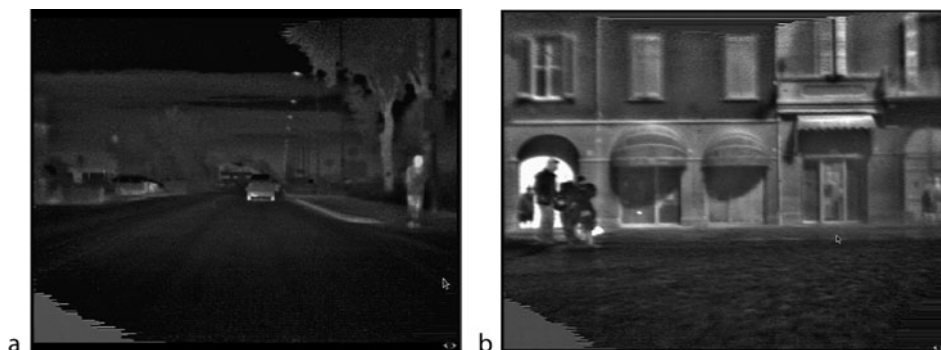
In the automotive environment, illumination can be barely controlled and therefore represents a major issue.

In fact, weather conditions, the different sun positions, and artificial light sources such as headlights or street lamps highly affect scene illumination. This is true for daylight or near infrared cameras, while in the case of far infrared cameras the problem arises only in extreme situations like direct sun framing or when light sources also produce thermal effects. Shadows represents a critical issue for image acquisition and processing; in fact, the simultaneous presence of shady and fully illuminated areas in the scene may lead to acquiring images in which shady areas are too dark or illuminated objects are too bright. Moreover, shadows represent a pattern that can interfere with the image processing systems based on pattern matching techniques. One case, in which shadows presence indirectly impacts on FIR domain as well, is due to thermal effect that light can have. In fact, sun or even artificial lights increases the temperature of objects exposed to lights creating *thermal shadows*; ● Fig. 38.5b shows this effect on the wall below the tents, which is colder than the other portions of the wall that are heated by the sun.

In addition, vehicle movements can lead to abrupt changes in illumination conditions. The worst situation happens when the sun is suddenly framed or exiting/entering a tunnel, making the entire image completely dark or white.

In such cases, cameras that have a fast Automatic Exposure Control (AEC) are recommended. AEC acts on both camera gain and control to compensate global illumination changes. Since a large gain value introduces noise at camera sensor level, it would be better to have a system that primarily acts on the shutter trying to maintain a low gain value; in addition, such a system can avoid to monitor the whole image reducing the area used for exposure control to the one actually processed. ● Figure 38.6 shows the result of an evolute exposure control algorithm that has been conceived to compute the most suitable exposure for the lower portion of the image, since the area of interest is the road and not the sky. In this case, the pedestrian can be recognized, while the computation of the exposure using also the upper portion of the image would have left the road completely dark. This requires a camera that features inputs for controlling gain and shutter like most IEEE1394 or IP-based cameras or a smart camera with some processing inside.

Smear effect The *smear effect* is another artifact degrading image quality for cameras in the visible domain: a strong light that directly hits the sensor in low illumination



■ Fig. 38.5

During winter, (a) FIR images allow an easy detection for pedestrians. In the summer, (b) the pedestrian at left is cooler than the background gate

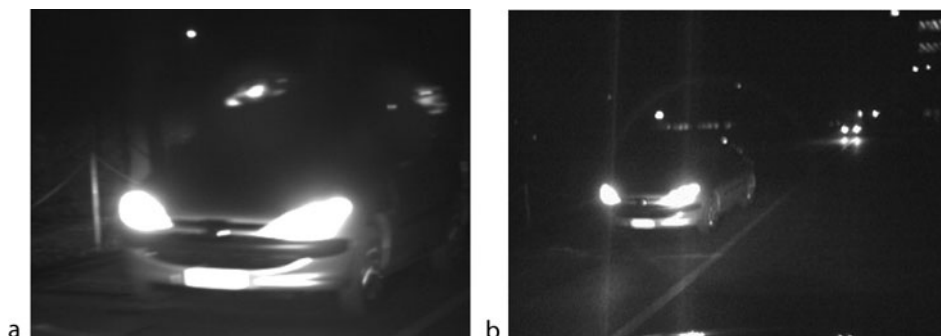


■ Fig. 38.6

Example of automatic exposure control obtained by defining a specific area (matched with the application) in which the contrast and brightness should assume optimal values

conditions produces bright artifacts, especially vertical bright lines (see ► Fig. 38.7a). This effect is typical for visible cameras and, in the automotive environment, can be easily caused by reflectors or other vehicles' headlights at night or inside tunnels. This effect can represent a source of noise for image processing leading to wrong results, that is, a lane markings detection system, that is typically based on the detection of bright lines on the road surface, can be fooled to interpret smear effects as lane markings.

Smear effect is caused by internal reflections inside the camera and lens system and is lower at the infrared wavelength. Therefore, near infrared cameras are less affected by this problem (see ► Fig. 38.7b) and can be evaluated as a replacement for standard daylight cameras in many situations.



■ Fig. 38.7

Smear effect in (a) visible cameras and (b) NIR devices



■ Fig. 38.8

Reflection of a far infrared radiation on a wet surface

Reflections and glares Reflection represents another source of problems for onboard systems.

The worst case is due to strong light reflections that dazzle the camera and lead to saturated images, but also weak reflections can create artifacts in acquired images. As an example, ► Fig. 38.8 shows how a wet asphalt road behaves as a mirror in the FIR domain and produces ghost pedestrians in the acquired image. In order to reduce reflections, a polarized lens can be used for cameras.

3 Conclusions

This chapter discussed the main issues that are very specific to the development of a vision system for a vehicle.

Installing vision systems on a vehicle lead to face problems that are generally more relaxed in the industrial scenarios like device positioning and wiring. A vision system is in fact composed by one or more camera devices and processing systems. At the same time, inside a vehicle, the vision system have often to be installed where there is not plenty of space and therefore requiring a careful selection of cameras, processing engines, and related communications.

Beside the installation problems also the functional issues have to be faced. A vision system inside a vehicle is generally moving and this leads to different problems that can affect the image processing like the incapability of controlling the light that can lead to critical situations, the oscillation, and vibrations that can affect inter or extra calibration, or the blurring in acquired images due to quick or repentine movements.

References

- Bombini L, Cerri P, Grisleri P, Scaffardi S, Zani P (2006) An evaluation of monocular image stabilization algorithms for automotive applications. In: Proceedings of the IEEE international conference on intelligent transportation systems 2006, Toronto, Canada, pp 1562–1567, Sept 2006
- Broggi A, Medici P, Porta PP (2007) StereoBox: a robust and efficient solution for automotive short range obstacle detection. *EURASIP J Embed Syst*. doi:10.1155/2007/70256
- Cardani B (2006) Optical image stabilization for digital cameras. *IEEE Control SystMag* 26(2):21–22
- Dickmans ED, Mysliwetz BD (1992) Recursive 3-D road and relative ego-state recognition. *IEEE Trans Patt Anal Mach Intell* 14:199–213
- Liang Y-M, Tyan H-R, Liao H-YM, Chen S-W (2003) Stabilizing image sequences taken by the camcorder mounted on a moving vehicle. In: Proceedings of the IEEE international conference on intelligent transportation systems 2003, Shangai, China, pp 90–95, Oct 2003
- Norén J (2008) Warning systems design in a glass cockpit environment. Thesis dissertation, University of Linkopings
- Schiehlen J, Dickmanns E (1994) Design and control of a camera platform for machine vision. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, Sendai, pp 2058–2063, Sept 1994

39 Camera Technologies

Paolo Grisleri

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Camera Definition</i>	1012
1.1	Camera Classification	1013
1.2	Sensor	1013
1.2.1	Dynamic Range	1014
1.3	Speed	1015
1.4	Image Format	1016
1.5	Exposure Parameters	1016
1.6	Optics	1017
1.7	Sensors	1017
1.8	Software	1019
1.9	Mechanical Issues	1019
1.10	Conclusions	1020

Abstract: This chapter starts with a theoretical definition of the sensor technology and describes the main parameters suitable to classify cameras. The section continues with a brief discussion on optics and sensors. A section then is dedicated to camera-specific software such as firmwares, API, or dedicated SDK. The last section describes mechanical issues specific to vehicular applications.

1 Camera Definition

Def. 1: An image (of the world) is a signal representing the 2D projection of one or more electromagnetic (EM) wave properties emitted or reflected by a 3D real-world scene, refracted through an optical system, and integrated over an appropriate amount of time.

The recorded properties can be intensity, phase, or frequency, but are usually the intensity of the EM at a certain frequency range, like in the case of visible images. But for far infrared images it can be the intensity of the emitted waves in a given range of wavelength. The frequency ranges can be one or more since in one image different information may be required, such as in the case of Bayer images, where adjacent pixels are dedicated to different wavelengths. Following this definition, the information itself can be indicated as “image,” not the data structure used to contain and manipulate this information.

If the image is focused on a sensor, it is necessary to fix an appropriate exposure time, depending on the sensor itself.

The optical system is tightly connected to the image taken since it captures the projection of the framed scene.

The quantized version (in space and in values) of the signal described in Def. 1 includes what is usually intended as a digital image: information that can be processed in numerical form and stored in different formats on common storage devices such as memory and disks.

When an image is acquired in a digital form, part of the information available in the real world collapses in the adaptation process to the new data structure. The biggest part of this loss is currently located in the dynamic range. Images are transferred using formats designed to cover dynamic ranges, allowed by current technology, but small compared to those available in real-world scenes, especially where illumination conditions are not controlled, such as outdoor.

After this clarification, a second definition can be introduced.

Def 2: A camera is a sensor suitable to capture a series of images in time.

This definition is general enough to include different camera types and components.

A chain formed by a lens group, a sensor, a processing system, and one interface for transmitting data can be reasonably considered as the simplest and most common case of camera.

A vision system may have more than one sensor, such as stereo cameras or omnidirectional cameras. Notice that in all cases the observed scene is the same. These kinds of cameras capture images from all their sensors, build a frame using this information, and send the frame to the processing unit in the usual way.

The time behavior of a camera can have several flavors. A camera can be controlled to take one or more images when a command is received; these modes are usually indicated as one shot or multiframe. A camera can be controlled to start capturing images when a command is received. The frame rate can be controlled internally by the camera or given as input from an external source.

A camera also provides access to modify the capture parameters or to read internal data; these parameters belong to one of the following categories:

- Image format: using this controls the data structure features produced by the camera; this category includes capture format, color coding, region of interest, pan, tilt, subsampling, binning, and decimation.
- Exposure: these controls allow to change the way the camera acquires the image. Examples of this category are exposure time (shutter), gain, white balance, sharpness, hue, saturation, and gamma.
- Trigger and IO: these define when and how the camera must capture frames; each camera has its own way to define how to manage external triggers and how to produce signals to control other devices whenever the capture of a frame starts or ends.
- Statistics: statistics on frames or the occupied bandwidth can be retrieved through appropriate interfaces.
- Processing: if present, these features allow to configure the onboard processing resources, such as DSP image processing.

1.1 Camera Classification

Cameras are classified depending on their features. This section explains each camera feature and the possible values where applicable, or the preferable features for using a camera in a specific field. Cameras are mainly classified depending on their sensor type; other features are directly related to the image dynamic range, speed, output image format, and adjustable exposure parameters. Other features more related to the application are triggers, IO functions, onboard processing functions, interface, probing, and statistics functions.

1.2 Sensor

Sensor is the main feature distinguishing one camera from another. The framed image is focused on the sensor plane through the optical system. A sensor is usually a rectangle of silicon where, using photolithographic techniques, a matrix of elements is created. Each element is made of a set of devices able to transform a fixed amount of incident radiation in a fixed amount of stored charge. Once the conversion process has been completed, another operation, called readout, has to be performed in order to transmit the image to the buffer. Sensors are characterized by their geometry, technology, spectral response, color filter array (for visible sensors), dynamic range, and type of shutter. Sensor geometry

is usually rectangular (though some exception exists, such as foveal sensors) and elements are placed in rows and columns. The amount of captured light increases with the cell size; for this reason bigger sensors offer a better signal to noise ratio (SNR) and dynamic range. Lenses can be created on the top of each element in order to increase the amount of captured light if the cell size is small. Color information is captured in different ways depending on the sensor target price. For low-cost sensors, the most common technique is the color filter array, where each element is covered by a different colored filter; filters are distributed on the sensor following a regular pattern, such as one of the Bayer pattern. Another technique to obtain color is to use a prism to separate the incident light into different color components such as red, green, and blue and use one monochrome sensor to capture each component separately. Recently, a new technology has been developed (Goi et al. 2006) which uses a three-layer CCD, where two of them are semitransparent; each layer detects one of the color components. This technology has a cost which is in between the color filter array and the 3CCD, but the images are much similar to the second one since they do not suffer from the aliasing problem due to the color reconstruction phase.

1.2.1 Dynamic Range

Current industrial cameras process signals using 12 bits, 14 in some cases. This depth is not enough to cover the everyday light range.

Luminance is defined as the amount of light intensity per unit area of light in a given direction. Luminance is measured in candela per square meter (cd/m^2). It is a measure of how much light is emitted from a surface through solid angle. When a human observes an emitting or reflecting surface, in a specific direction, his eye defines a solid angle. Luminance is the amount of light emitted by the surface and entering through the solid angle in that direction.

Typical luminance ranges from 0.1 cd/m^2 during night scenes to more than 50.000 cd/m^2 .

The dynamic range of a scene is defined as:

$$DR = 20 * \log L_M / L_m$$

where L_M is the maximum luminance in the scene in $[cd/m^2]$, L_m is the minimum luminance in the scene in $[cd/m^2]$.

In high contrast scenes, DR can be higher than 110 dB. The human decreases its sensitivity with luminance and can perceive a huge range of luminance: between 3×10^{-6} and $1 \times 10^9 \text{ cd/m}^2$. There is a problem in both capturing and representing this data. It is currently not possible to produce sensors able to cover such a huge dynamic range. In any case, it would be necessary to change the way images are represented in memory for processing this kind of data. Currently, images of up to 14 bit depth are used in professional photography to represent 1 pixel.

To cover the same range of the human eye with a linear scale, 50 bits would be necessary. This number can be compressed to not less than 32 bits using an appropriate

logarithmic transfer function. This is exactly what is currently done to adapt the sensor range to the output, file, or display, which is more dependent on historical background. The most part of available displays and file formats support up to 8 bits per channel. Some proprietary format for professional use can store up to 14 bits.

“High dynamic range,” collects a set of techniques to generate images encapsulating a dynamic range higher than usual. These techniques are basically two:

1. Take more images with different exposures
 - (a) Using the same sensor shooting at different times
 - (b) Using sensors placed in slightly different positions shooting at the same time
2. Using special HDR sensors

The first technique is used after the capture to merge the images in one single HDR image using appropriate algorithms (Gortler and Myszkowski 2005; Reinhard et al. 2005). The second technique requires special hardware. These techniques are today commonly used for photography but they are suitable for still scenes framed from a fixed camera.

In both cases, special displays and file formats such as 3FR, NEF, CR2, or DNG are needed to show and store the images.

Automotive applications have more demanding constraints. The camera and background are moving, as well as the framed objects. In this scenario, taking images in different times or from different geometrical positions leads to unacceptable artifacts. Sensors like the MT9V022 from Micron are available on the market and produce HDR images by compressing a 12 bit range to 8 bits. The transfer function between light intensity and video output level, can be selected between linear, to cover a dynamic range is 55 dB, or piecewise linear to reach 100 dB. Other sensors can reach 120 dB using a logarithmic transfer function. The VL5510 CMOS sensor from STMicroelectronics is specific for automotive applications. CMOS sensors allow the pixel response to be different in different areas of the image, useful especially in hard illumination conditions such as entering or exiting from tunnels and with strong shadows. CMOS sensors featuring back-thinning and rear illumination processes are currently under development and beginning to appear on the market.

Some cameras allow to periodically transfer a frame series taken in a range of different exposure conditions (bracketing). Each frame series can be quickly analyzed and selected for further processing, a frame containing a correctly exposed selected region of interest. When the framed scene is still slowly moving, the images captured with the bracketing technique can be joined into a unique HDR image.

1.3 Speed

The maximum number of frames a camera can produce in a fixed amount of time is a critical factor for some applications. Usually, this number is given by a trade-off between the bandwidth generated by the camera – due to exposure time, frame rate, frame size, depth, and encoding – and the available bandwidth on the communication interface with

the storage or processing device. The application itself typically determines the maximum required speed, for example, in prototype automotive application such as pedestrian detection or vehicle detection systems generally work at 10 fps, while filming fast events such as crash tests 1,000 fps rates are pretty common. Budget constraints can be overridden in some cases using camera features. For example, the bandwidth can be reduced sending the raw (Bayer) frame on the destination and performing color reconstruction on the processing system.

1.4 Image Format

Digital cameras produce a data stream in a format specified by the manufacturer. The stream can be modified, through either a dedicated API or controls integrated on the camera itself. Parameters that can be modified include the following:

- Image geometry: these parameters define image geometry (width and height); a region of interest can be specified to capture only a subarea inside the image geometry; when using region of interest the frame rate can be increased since the amount of data to be transmitted is lesser.
- Color depth: this parameter controls the image color format, and onboard color reconstruction can be assigned; in this way a raw image can be selected to save bandwidth between the camera and processing unit or a full RGB8 format can be selected to save CPU power.
- Binning, decimation, subsampling: some cameras allow to return different types of subsamples of the imager data. Binning will take 1 pixel for every two for rows and columns; decimation 1 pixel every ten for each row and for each column. Other cameras subsample color reconstruction from a Bayer image: in this case a full RGB image with half width and half height of the original image is produced; no aliasing (Farsiu et al. 2006) is present since each pixel of the output image covers one entire pattern in the sensor.

1.5 Exposure Parameters

This set of features allows to change the characteristics of the acquired images. Some of these features have automatic controls to continuously adjust the value in order to track a certain target using a specific algorithm. For example, on most industrial cameras, enabling the AutoExposure control will enable the camera to change continuously its gain and exposure values to maintain a certain target exposure; also, cameras with more advanced electronics allow to choose between different autoexposure algorithms and to limit shutter and gain variations to a fixed range. Exposure is controlled by the following parameters:

- Shutter: changes the acquisition time for each frame.
- Gain: changes the pixel gain value over the entire image, also called contrast.
- Brightness: this control is an offset over the whole image.
- Auto exposure: changes shutter and gain using a specified algorithm to follow a specified target.
- White balance: changes the separate gain over two colors to compensate different illumination types.
- Hue: changes the mapping between colors.
- Saturation: increases the color values.
- Gamma: changes the nonlinearity used to map the acquired values to the output values.

1.6 Optics

Lenses are the part of the camera which gather and bend the light from the scene to focus it on the sensor. To obtain a better image, every optic is usually made by a group of lenses, reducing problems such as vignetting (Zheng et al. 2006) and chromatic aberration (Kang 2007); some lenses can be controlled to move, with the aim of removing vibrations. Optics integrates two controls: focus and iris. The first moves internal lenses to change the focus plane from near to far; the second controls the amount of light that hits the sensor as well as the depth of field. Some lenses integrate motors to change these values; autofocus and autoiris signals can be sent directly by the camera if supported.

1.7 Sensors

The translation from incident radiation coming from the optical system to an electrical signal is performed by the sensor. Sensors are classified depending on their technology, geometry, size, and spectral response. The two main technology branches available today for the visible domain are near infrared (NIR) CCD and CMOS. The charge-coupled device (CCD) development started in late the 1960s thanks to the relatively simple technology level necessary to produce these devices. It is currently widespread since during the long development time this technology has been pushed to its limit. High-quality applications take advantage of the CCD longevity thanks to the high signal to noise ratio (SNR) and the high uniformity of the image. The biggest part of the sensor space can be dedicated to photo-detectors. This feature allows to capture more incident light compared to CMOS; however, the hardware needed to build the image have to be placed outside the sensors increasing camera cost and complexity. One point against the CCD technology is the blooming effect: framing high light sources in dark scenes results in horizontal or vertical straight lines, traversing all sensors depending on the mounting orientation. These artifacts may disturb vision algorithms. This is due to the reading mechanism: the charge is transferred across the adjacent elements of a row and then

through a column up to the sensor border and there converted into a voltage. Today, front-illuminated CCDs are very common, easy to produce, and thus relatively low cost. Photo-detectors are illuminated by incident light but part of this light is reflected or absorbed. This reduces the quantum efficiency: the amount of light converted in charge. Back-illuminated CCDs are obtained by removing most of the silicon bulk from a front-illuminated sensor and mounting it upside down. In this way the incident light traverses the remaining thinned bulk and hits the sensible photo-detector area without impact on the gate or the insulator. The quantum efficiency of this technique is 50% greater than the front-illuminated sensor, making these devices very attractive for ADAS and low-light applications. Photons penetrate more deeply in silicon as their wavelength grows. Fabricating sensors using high resistivity, deep depletion silicon increases the NIR sensitivity; however they must be employed only when the NIR range is effectively required by the application since these kinds of devices are affected by high dark current effects (this means that the image when a black scene is framed is not completely black).

With CMOS technology, the sensor area available to capture light is reduced since the circuitry to convert charge into voltage and for signal conditioning is distributed nearby each pixel. CMOS technology was born in the late 1970s and developed during the 1990s. Low cost and performance are the key features driving the success of CMOS sensors over CCDs in these last years.

CMOS is not affected by blooming, and has a lower power consumption compared with CCD since the readout operations involve each pixel only once. One of the main CMOS issue is the noise due to the tight integration of devices. It is made of two components: a fixed pattern noise which can be removed subtracting a premeasured pattern, and a random noise. Image quality is also affected by the transistor length reduction: linearity and dynamic range.

High dynamic range is primary for automotive applications since outdoor scenes may have a huge range of luminance values. Under bad illumination conditions, such as when the sun is framed or during night driving, a linear range mapped over 8 bits or even 16 bits is too low to represent the scene. Examples of these conditions include abrupt brightness changes or when the sun hits directly the sensor. In these cases the processing would be often impossible even for a human. Fast automatic exposure controls or HDR techniques can be used in these cases.

Today more than 40 vendors supply imaging sensors. Camera vendors mostly use products from Kodak and Sony. However, some specific markets such as the automotive require special features that are being fulfilled by vendors targeting specific needs.

CMOS sensors are based on active-pixel arrays: Each pixel is made by a photo-site and readout logic. For this reason the fill factor cannot reach 100%. Although these devices have lower performance than CCDs in terms of image quality, the lower cost and complexity for cameras can justify their spread in low-cost, high-volume applications such as mobile phones. In the last years, companies like DALSA and Canon pushed this technology producing sensors and cameras with performances comparable to those of CCD. In most specific markets such as automotive, where high dynamic ranges and high sensitivity in low-illumination conditions are needed, CMOS can gain a dominant position over CCD.

1.8 Software

One of the key factors that contribute to determine the success of a camera is the availability of an SDK to develop custom applications. Since every camera has its own peculiarities, it is up to the manufacturer to supply this kind of SDK over different platforms. Sometimes several manufacturers join in consortiums to establish new standard interface for cameras and support the same hardware interface and protocol. Examples of these protocols are FireWire + DCAM and GigE; the first is based on the IEEE1394 hardware and describes the exchange of video and control data with IEEE 1394 cameras; the second is more recent and uses the Gigabit Ethernet hardware and describes how to transmit uncompressed images and control data.

Low-level capture libraries provide C/C++ API to efficiently manage every aspect of the acquisition, such as image format, camera features, triggers and I/O, statistics, and camera information. Users can write their own applications including the manufacturer header file and linking the library provided with the SDK.

Software libraries capable of capturing images and controlling different types of cameras are available (Bertozzi et al. 2008). These kinds of software allow to write processing software independently of the hardware. In this way, whenever a camera is no longer available from the manufacturer, a replacement from another manufacturer can be used without changing the application software. This kind of solution requires the application to be independent of camera parameters.

1.9 Mechanical Issues

Several constraints have to be taken into account when installing a camera on a vehicle. Cameras need to be placed in different positions in order to obtain the best detection performance for a specific system. Some positions were proven to be more effective. Usually the optical axis is oriented in the forward direction, this being the direction where the driver's sight is concentrated most of the time. Looking in this direction, obstacles, pedestrians, traffic signs, and lane markings can be detected. Such cameras can be mounted internally behind the front windshield, nearby the internal rearview mirror. This position has the advantage of being cleaned by wipers. Images taken from the described position are strongly blurred by rain drops and dirt on the near windshield surface. Forward-looking cameras can also be placed outside in the front of the vehicle. However this position is not so efficient since the camera height is usually limited by the hood: this low angle of observation usually jeopardizes the vision system capacity. Cameras with wide aspect ratio sensors are preferable for this orientation in order to avoid framing the hood and the sky. Reflections are another important factor: mounting a camera behind the windshield may lead to artifacts in the images, especially during sunny days.

Another important position is into the external rear view mirrors, looking backward, to detect overtaking vehicles and road lane markings. Other positions can be in the back of the car, looking at the maneuver area.

1.10 Conclusions

This chapter started with a theoretical definition of cameras in a general form. Parameters such as sensor, dynamic range, speed, image format, and exposure parameters have been presented to perform a camera classification depending on the application needs. A section briefly describes optics and the main issues connected to them are reported, followed by a deeper discussion on the sensor technology clarifying the main differences between CCD and CMOS. In recent cameras, more features are always available on a software interface, and thus a section has been dedicated to this topic. A section dedicated to mechanical issues related to the automotive field closes the chapter.

References

- Bertozzi M, Bombini L, Broggi A, Cerri P, Grisleri P, Zani P (2008) GOLD: a complete framework for developing artificial vision applications for intelligent vehicles. *IEEE Intell Syst* 23(1):69–71
- Farsiu S, Elad M, Milanfar P (2006) Multiframe demosaicing and super-resolution of color images. *IEEE Trans Image Process* 15(1):141–159
- Goi HK, Giesbrecht JL, Barfoot TD, Francis BA (2006) Electronic imaging: digital photography II. In: *SPIE-proceedings*, 22(6069), 2006
- Gortler SJ, Myszkowski K (2005) A collision mitigation system using laser scanner and stereovision fusion and its assessment. In: *Real-time high dynamic range texture mapping*, pp 313–320, June 2005
- Kang SB (2007) Automatic removal of chromatic aberration from a single image. In: *IEEE conference on computer vision and pattern recognition, CVPR '07*, Redmond, pp 1–8, June 2007
- Reinhard E, Ward G, Pattanaik S, Debevec P (2005) *High dynamic range imaging: acquisition, display and image-based lighting*. Morgan Kaufmann Publishers, Amsterdam/Boston. ISBN 978-0-12-585263-0
- Zheng Y, Lin S, Kang SB (2006) Single-image vignetting correction. In: *IEEE conference on computer vision and pattern recognition 2006*, New York, June 2006

40 Perception Tasks: Lane Detection

Luca Mazzei · Paolo Zani

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Introduction</i>	1022
2	<i>Lane Detection Requirements</i>	1023
3	<i>A Lane Detection Algorithm</i>	1023
3.1	Low-Level Processing	1024
3.2	Lane Marking Extraction	1025
3.3	Tracking and Generation	1025
3.4	Expansion and Validation	1027
4	<i>Results</i>	1028
5	<i>Conclusions</i>	1030

Abstract: The localization of painted road markings is a key aspect of environment reconstruction in urban, rural, and highway areas, allowing a precise definition of the safely drivable area in front of the vehicle.

Lane detection algorithms are largely exploited by active safety systems in the automotive field, with the aim of warning the driver against unintended road departures, but are also essential in fully autonomous vehicles, since they complement the data coming from other sources, like digital maps, making it possible to navigate precisely even in complex scenarios.

This chapter introduces potential approaches and requirements for lane detection and describes in detail one of such algorithms and its results.

1 Introduction

Lane markings are a very common, relatively inexpensive infrastructure designed to provide visual cues to the driver, in order to ease the task of driving, making it both safer and more time efficient. Lane markings availability becomes of great importance while driving when visibility is reduced, as it happens in case of fog or heavy rain, and when blinding occurs, for example, because of the headlights of oncoming traffic; the number, color, and pattern of the lane boundaries encode information about which maneuvers can be performed, such as overtaking or parking, or whether a lane is reserved to certain vehicle types (e.g., busses, taxis, and car-pools).

The widespread availability of lane markings, the amount of information they convey, and the fact that they are designed for easy visual identification makes them a promising candidate for automatic recognition; however, other solutions to the problem of defining a drivable path have been proposed, most notably magnetic guidance and digital maps.

Magnetic guidance systems exploit magnets buried in the road pavement and a detector installed on the vehicle to supply a reference trajectory, thus allowing accurate positioning within the lane, with errors in the range of only a few centimeters; the magnets polarity alternance can also be exploited to encode additional information about the upcoming road segment (Zhang 1991; Bin Zhang and Parsons 1994). While the performance is generally good in all weather conditions, to date the extra infrastructure needed for the system to work is not available to any significant extent, as just some pilot projects have been set up. Commercial exploitation of these systems has mostly been targeted at public transportation, where autonomously guided busses mostly travel in dedicated lanes, under the supervision of a human driver (Shladover 2007).

The use of digital maps has seen a significant development in recent years, with an ever increasing level of detail. Such maps, however, are hard to maintain up to date with respect to the physical status of the roads, and even small discrepancies can render them useless; moreover, were dense maps widely available, the problem of self-localization within them would still exist. Consumer GPS solutions are not typically accurate enough, especially when the satellites signal is subject to outage or multipath effects, like in city downtowns; unfortunately, these are usually the areas where the road layout gets more complex, making the environment reconstruction task very challenging.

With so many emerging technologies available, it is important to note how experiences like the 2007 DARPA Urban Challenge (Buehler et al. 2009) have shown that the fundamental requirement for safe operation of intelligent and autonomous vehicles, especially when sharing the road with regular traffic, is the capability of extracting updated information from the surroundings in real-time. The exploiting of the same infrastructure used by human drivers makes the coexistence easier, by reducing discrepancies in perceived information to a minimum.

2 Lane Detection Requirements

The task of identifying painted road markings consists in processing the available sensor data, possibly merging it with existing information on the environment, in order to determine whether a lane marking exists in the vehicle surroundings, and to define the resulting lane properties:

- Geometry – the physical layout of the drivable surface (like width and curvature)
- Topology – how lanes start, stop, merge, split, or intersect
- Semantics – the driving rules associated with the marking, such as the driving direction and whether lane change maneuvers are allowed or not

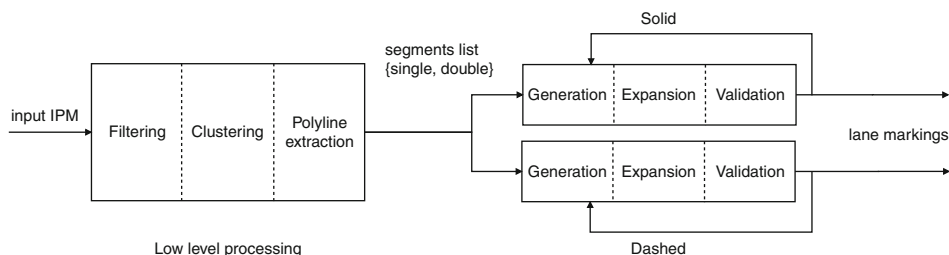
Depending on the intended application field, the level of required perception detail can vary greatly: *Lane Departure Warning* (LDW) systems usually assume an highway-like scenario, with a driver controlling the vehicle most of the time, and only need to detect the boundaries of the lane the vehicle is already traveling in; conversely, fully autonomous vehicles need to achieve a better understanding of the surroundings, in order to perform lane changes, negotiate intersections, and more in general, obey traffic regulations.

Being less demanding, LDW systems tend also to have lower hardware requirements, both in terms of sensors availability and computational power: a typical setup includes a forward looking camera mounted behind the rear-view mirror plus a GPS and/or *Inertial Measurement Unit* (IMU), thus allowing to build a coarse model of the lane, and to predict the vehicle trajectory in the immediate future.

An autonomous vehicle typically features more expensive GPS and IMU solutions (such as differential GPS) and extra sensors, like LIDAR and RADAR units or multiple cameras, in order to obtain a larger field of view and to suppress false detections caused by the presence of other vehicles and roadside elements like guardrails, poles, trees, and buildings; the processing hardware also needs to be scaled accordingly.

3 A Lane Detection Algorithm

This section describes a lane markings detection algorithm designed to run on an embedded platform connected to a single camera. While this hardware setup is typical of simpler applications, no a priori hypothesis on the number of markings or their



■ Fig. 40.1
System architecture

geometry has been enforced within the algorithm, as it is otherwise often the case. This choice allows a much broader application scope, including fully autonomous driving. To accomplish the more demanding tasks, additional sensors can be used:

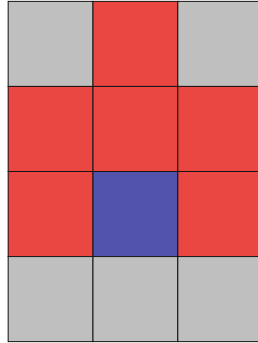
- An IMU can be exploited to estimate the instant extrinsic camera calibration parameters.
- LIDARs or stereo vision to detect obstacles, which can then be turned into a mask of invalid points in the image to process.

The overall system architecture is presented in [Fig. 40.1](#): first an Inverse Perspective Mapping (IPM) transformation (Mallot et al. 1991; Bertozzi et al. 1998) is applied to the frame grabbed by the camera, then a low-level filtering is performed to highlight the Dark-Light-Dark (Nedevschi et al. 2006) (DLD) patterns of the image; the resulting points are grouped together, and clusters are finally approximated by continuous piecewise-linear functions. Once the low-level processing is over, the resulting segment lists are compared to existing lane markings in a tracking stage ([Fig. 40.1](#)), which produces a set of candidate lane markings; moreover, non-tracked segments are also analyzed to extract additional candidates. An expansion step is then performed, in order to join in any pertinent non-connected component still present. Finally, each candidate is assigned a score which is tested against an acceptance threshold to produce the end result. The lane markings detection is carried out for solid and dashed markings, using slightly different algorithms, and the whole procedure is performed two times, one for single and one for double lines.

3.1 Low-Level Processing

The low-level processing stage derives from the one used during the 2007 DARPA Urban Challenge, described in (Broggi et al. 2010). First DLD and DLDLD transitions (with the former corresponding to single lane marking, and the latter to double ones) are extracted from the IPM and stored in separate buffers; this operation is fast since the filtering kernel is of constant size (5 and 11 pixels respectively).

After pattern extraction, a binarization step is performed: this process employs a variable threshold proportional to the average luminance of the region, over a window



■ Fig. 40.2

Clustering algorithm expansion mask: in dark grey, the reference pixel; in grey, the candidates for expansion. The topmost candidate helps to reduce the likelihood of cluster fragmentation due to non-continuous groups of points (e.g., because of dirty or faded lane markings)

of size 32×1 pixels; this helps to reduce the effects of shadows cast by vehicles and roadside elements, like buildings, trees, and guardrails. The resulting pixels are then grouped together by a clustering algorithm which uses the expansion mask illustrated in [Fig. 40.2](#), with the processing taking place starting from the bottom of the image. The various groups of points are then approximated using piecewise-linear functions, so that each node on the polyline corresponds to an element of the cluster, and the maximum distance between any pixel within the label and the closest segment is below a given threshold. While different solutions to this problem exist (like the one described in (Hakimi and Schmeichel 1991)), the one described in this chapter has proven to be good enough to deal with the data produced by the preprocessing stage.

3.2 Lane Marking Extraction

Once the set of polylines has been extracted from the image the actual lane markings extraction can take place. This process consists of three fundamental stages:

- Tracking of markings detected in previous frames
- Generation of new candidates
- Expansion of the whole set of candidates (new and tracked)

That are carried out in that order both for solid and dashed markings, and are detailed in the following.

3.3 Tracking and Generation

Candidate polylines are matched against existing lane markings, and only those resulting close enough can be considered a valid correspondence.

The fundamental issue to solve when performing this task is to define a distance function: a number of well-known approaches to this problem exist, like the Hausdorff (Hangouet 1995), Fréchet (Frchet 1906), and minimum Euclidean distances (Peuquet 1992), but while they exhibit some interesting mathematical properties, sometimes they can produce counter-intuitive results; instead, area-based algorithms (McMaster 1986) seem to be more appropriate in this kind of applications. Building on this idea, the distance between two polylines a and b becomes

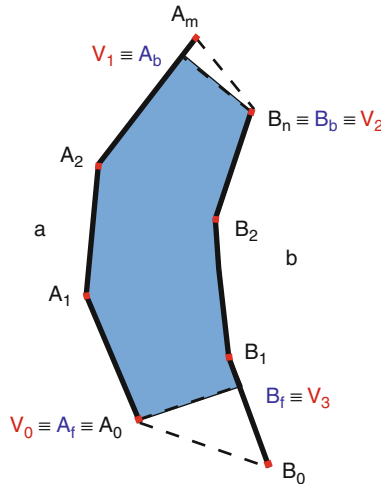
$$d(a, b) \stackrel{\text{def}}{=} \max \left(\frac{2 \times \text{area}(a, b)}{\text{length}(a)_{[V_0, V_1]} + \text{length}(b)_{[V_3, V_2]}}, d_{\min}(a, b) \right) \quad (40.1)$$

where $\text{length}(p)_{[V_m, V_n]}$ denotes the length of the polyline portion delimited by the points V_m and V_n , $\text{area}(a, b)$ is the area between the polylines (marked in light blue in Fig. 40.3), and $d_{\min}(a, b)$ is the minimum distance between a and b .

When performing tracking, each solid lane marking $ts_i \in \{ts_0 \dots ts_k\}$ identified at time $T - 1$ is used to compute the score

$$\text{score}(ts_i, cs_j) \stackrel{\text{def}}{=} \frac{d(ts_i, cs_j)}{\text{length}(cs_j)}, j \in \{0 \dots h - 1\} \quad (40.2)$$

matching it against the candidates $\{cs_0 \dots cs_{h-1}\}$ generated by the low-level processing stage at time T ; the candidate obtaining the lowest value is selected, and used in the following expansion stage.



■ Fig. 40.3

Distance between polylines. In gray, the overlapping area; in gray points, boundary vertexes; in black, the projections of one polyline ends onto the other

Dashed lane marking tracking uses a different comparison criterion, since the length of candidate dashes $\{cd_0 \dots cd_{k-1}\}$ is usually too small to be reliable:

$$score(td_i, cd_j) \stackrel{\text{def}}{=} d(td_i, cd_j)^2 + x_j^2, j \in \{0 \dots k-1\} \quad (40.3)$$

with x_j being the x component of the first vertex of cd_j . This heuristic allows to obtain as a winning candidate a dash that is near to the lane marking to track while also being as close as possible to the camera, that is, in the IPM region that is more likely to produce accurate results.

After the first dash has been determined, it is used as a starting point to join in any other dash close to the originating lane marking (that is, td_i), iteratively building a new polyline cp . The criterion adopted to perform this task aims at isolating candidates being close both to cp and to td_i , sorting dashes according to the following comparison rule:

$$\min(cd_h, cd_k) \stackrel{\text{def}}{=} \begin{cases} cd_h & \text{if } d(cd_h, td_i) < th \text{ and } d(cd_k, td_i) > th \\ cd_k & \text{if } d(cd_h, td_i) > th \text{ and } d(cd_k, td_i) < th \\ \arg \min (d(cd, cp) + d(cd, td_i)) & \text{otherwise} \\ cd \in \{cd_h, cd_k\} \end{cases} \quad (40.4)$$

Each time a dash is added to cp the remaining ones are sorted again using [Eq. 40.4](#), until no close dashes are left. To further improve the robustness of this step, dashes are joined only if they satisfy the constraints illustrated in [Fig. 40.4](#), and further explained in [Sect. 3.4](#).

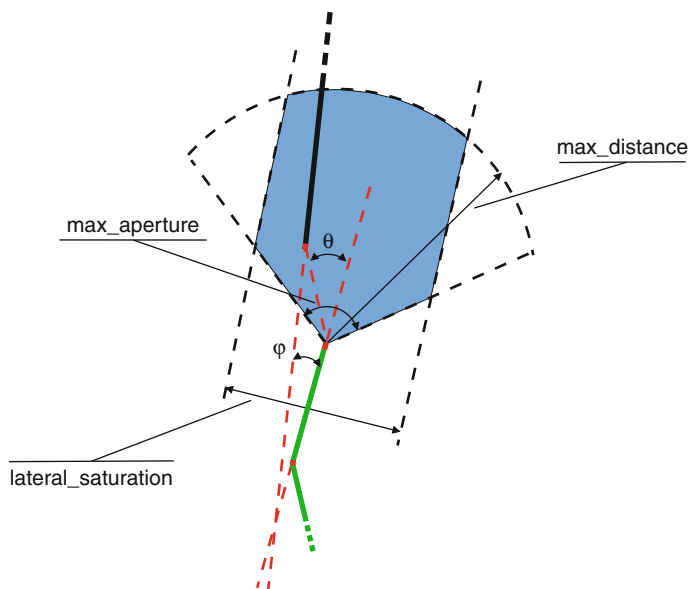
3.4 Expansion and Validation

Tracked and non-tracked polylines are iteratively analyzed to determine whether any other compatible candidate exists; if it is found, its points are merged in, and the search continues using the resulting polyline as the new reference. For both solid and dashed lane markings, a common condition for inclusion is that the first vertex of the candidate polyline must be close to the last vertex of the reference one, as it is illustrated in [Fig. 40.4](#); moreover, the orientation of the end segments must be similar (that is, the angle φ in [Fig. 40.4](#) must be small), and the connection angle (θ in [Fig. 40.4](#)) must also be small. Dashed markings bear the additional constraint that dash lengths and pauses between dashes must have a similar length.

When the search is over, a score is assigned to the resulting polyline p , to determine whether it should be accepted as a valid lane marking or not. The value is computed as

$$score(p) \stackrel{\text{def}}{=} \frac{length(p)^2}{d((0,0), p)^2} \quad (40.5)$$

and if p has been successfully tracked, the old score is added to the current. Using this approach means that lines starting far away from the vehicle are considered valid only if



■ Fig. 40.4

Compatibility test. In *green*, reference marking, in *black*, candidate polyline to join. The candidate is considered for inclusion only if its first vertex falls inside the blue area (determined by the parameters *max_distance*, *max_aperture* and *lateral_saturation*), and the angles φ and θ are small enough

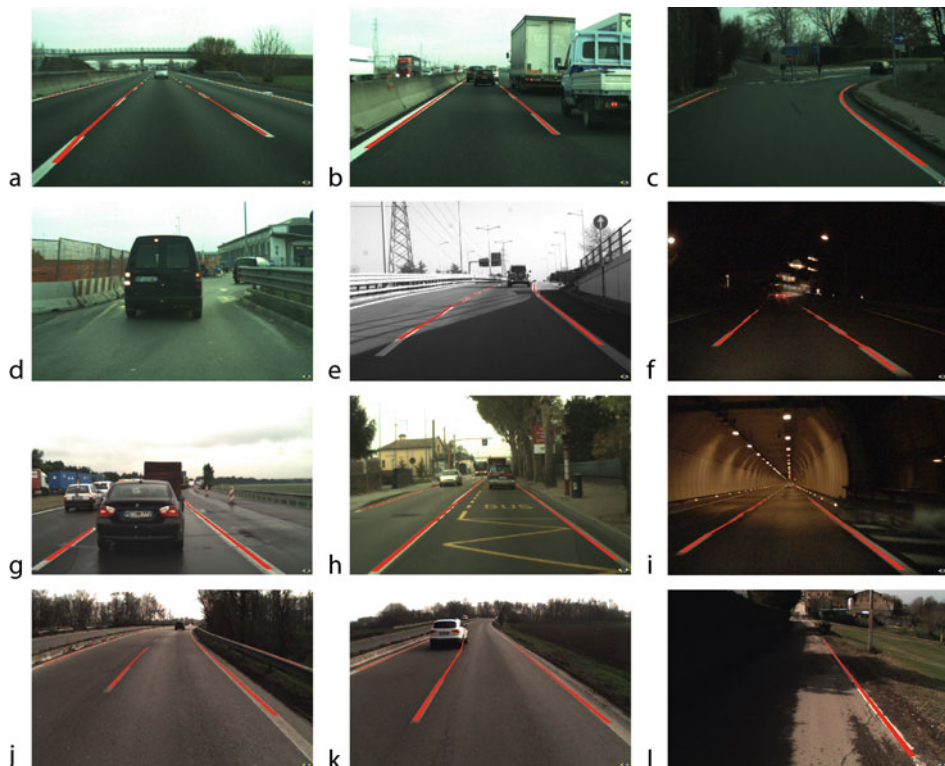
they are consistently detected over a high number of frames, while close, long lines are more easily accepted.

Tracked lane markings are not discarded immediately in case of a missed detection; instead, they are kept as “ghosts” for up to five frames (the value has been determined empirically).

4 Results

► *Figure 40.5* contains some samples from different image sequences, along with lane detections results.

In ideal working conditions, the system has proven to be highly reliable (► *Fig. 40.5a*), even in case of heavy traffic (► *Fig. 40.5b*), and tracking is very stable. Since no predefined model of the road is enforced atypical geometries can be handled as well, like in ► *Fig. 40.5c*, where lane markings leading to an intersection at the end of a downhill road are correctly detected. The adopted score criterion shows its effectiveness in ► *Fig. 40.5d*: no false detection is introduced, despite the number of objects present in the scene, including a vehicle right in front of the camera, the striped barriers on the left, and the guardrail on the right.



■ Fig. 40.5

Some sample outputs in different situations: (a) Highway scenario; (b) Heavy traffic; (c) A downhill intersection; (d) Construction area, with no false detections; (e) Uphill motorway with shadows; (f) Nighttime highway under heavy rain; (g) Queued vehicles; (h) Urban environment; (i) Entering an highway tunnel at night; (j) False detection due to a guardrail; (k) The side of a car interpreted as a dash; (l) Country road with strong shadows

► [Figure 40.5e](#) and ► [Fig. 40.5f](#) contain the results obtained in two challenging situations, namely on a motorway with a lot of shadows and under heavy rain at night: in both cases the algorithm correctly detects the road markings.

Other challenging situations are shown in following scenarios. ► [Figure 40.5g](#) shows a vehicles queue; while ► [Fig. 40.5h](#) shows an urban road environment with some other lanes and a bus stop. ► [Figure 40.5i](#) shows a tunnel entrance, a challenging scenario because of the abrupt illumination changing. While in previous critical situations the system has proven to be reliable in other frames small errors occurred.

► [Figures 40.5j](#) and [k](#) show examples of false detection related to guardrail and rear side of a preceding vehicle that are detected as a line marker. Geometry and color of this object in the image misled the algorithm. ► [Figure 40.5l](#) shows a country road with high shadow, lane detection system is able to detect only the left road lane.

5 Conclusions

The information about the road boundaries is of paramount importance to both human and autonomous drivers, and as such it must be encoded effectively. While different technologies can suit this need, depending on a number of constraints, painted lane markings still seem to be the most practical and cost-effective candidate to simultaneously allow automatic and visual detection, while conveying geometrical, topological, and semantical information on the road.

Depending on the field of application, the automated detection of lane markings can be carried out exploiting data coming from different sensors, possibly fused together. Warning-only systems can be effective even with loosely calibrated sensors, and as such are easier to integrate in mainstream vehicles; conversely, fully autonomous navigation requires a much more precise detection of the lanes geometry, which can be obtained through the use of a number of different and accurately calibrated sensors, and computationally intensive algorithms.

The lane detection algorithm described in this chapter has been designed from the ground up to scale well as more sensors are available, so that it can be deployed in a variety of scenarios. To assess its performance in real-world conditions, it has been tested during the VIAC expedition (VisLab 2010; Bertozzi et al. 2010), during which it has successfully negotiated a wide variety of environments, ranging from eastern Europe to China, all with unique weather and road conditions, demonstrating the robustness of such an approach.

References

- Bertozzi M, Broggi A, Fascioli A (1998) Stereo inverse perspective mapping: theory and applications. *Image Vis Comput J* 8(16):585–590
- Bertozzi M, Bombini L, Broggi A, Buzzoni M, Cardarelli E, Cattani S, Cerri P, Debattisti S, Fedriga RI, Felisa M, Gatti L, Giacomazzo A, Grisleri P, Laghi MC, Mazzei L, Medici P, Panciroli M, Porta PP, Zani P (2010) The VisLab intercontinental autonomous challenge: 13,000 km, 3 months, no driver. In: *Proceedings of the 17th World congress on ITS*, Busan, South Korea, Oct 2010
- Bin Zhang W, Parsons R (1994) Intelligent roadway reference system for vehicle lateral guidance and control, United States Patent 5347456, September 1994
- Broggi A, Cappalunga A, Caraffi C, Cattani S, Ghidoni S, Grisleri P, Porta PP, Posterli M, Zani P (2010) TerraMax vision at the urban challenge 2007. *IEEE Trans Intell Trans Syst* 11(1):194–205
- Buehler M, Iagnemma K, Singh S (2009) The DARPA urban challenge: autonomous vehicles in city traffic. In: *The proceedings of the DARPA Urban Challenge*, Seattle, 2009
- Frchet MM (1906) Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884–1940)*, 22(1):132–136
- Hakimi SL, Schmeichel EF (1991) Fitting polygonal functions to a set of points in the plane. *CVGIP: Graph Models Image Process* 53(2):132–136
- Hangouet JF (1995) Computation of the Hausdorff distance between plane vector polylines. In: *Proceedings of the twelfth international symposium on computer-assisted cartography*, vol 4, Charlotte, North Carolina, USA, pp 1–10
- Mallot HA, Bülthoff HH, Little JJ, Bohrer S (1991) Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biol Cyberne* 64:177–185
- McMaster RB (1986) A statistical analysis of mathematical measures for linear simplification. *Cartograph Geograph Inform Sci* 13(2):103–116
- Nedevschi S, Oniga F, Danescu R, Graf T, Schmidt R (2006) Increased accuracy stereo approach for 3D

- lane detection. In: IEEE intelligent vehicles symposium, Tokyo, Japan, pp 42–49, June 2006
- Peuquet DJ (1992) An algorithm for calculating minimum euclidean distance between two geographic features. *J Comput Geosci* 18(8):989–1001
- Shladover SE (2007) Lane assist systems for bus rapid transit, Volume I: Technology Assessment. California path research report UCB-ITS-PRR-2007-21
- VisLab (2010) VIAC web site. <http://viac.vislab.it>
- Zhang W-B (1991) A roadway information system for vehicle guidance/control. In: Vehicle navigation and information systems conference 1991, vol 2, Dearborn, pp 1111–1116, Oct 1991

41 Perception Tasks: Obstacle Detection

Stefano Debattisti


Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Introduction</i>	1034
2	<i>Sensors for Obstacle Detection</i>	1036
2.1	Proprioceptive and Exteroceptive Sensors	1036
2.2	Active and Passive Sensors	1036
3	<i>Obstacle Detection Methods</i>	1038
3.1	Active Sensor Use	1038
3.2	Vision	1038
3.3	Multi-sensor Fusion	1040
4	<i>Conclusion</i>	1041

Abstract: Obstacle detection is a widely studied field in the automotive industry because of the great importance it assumes in all systems that provide autonomous navigation of vehicles in an environment.

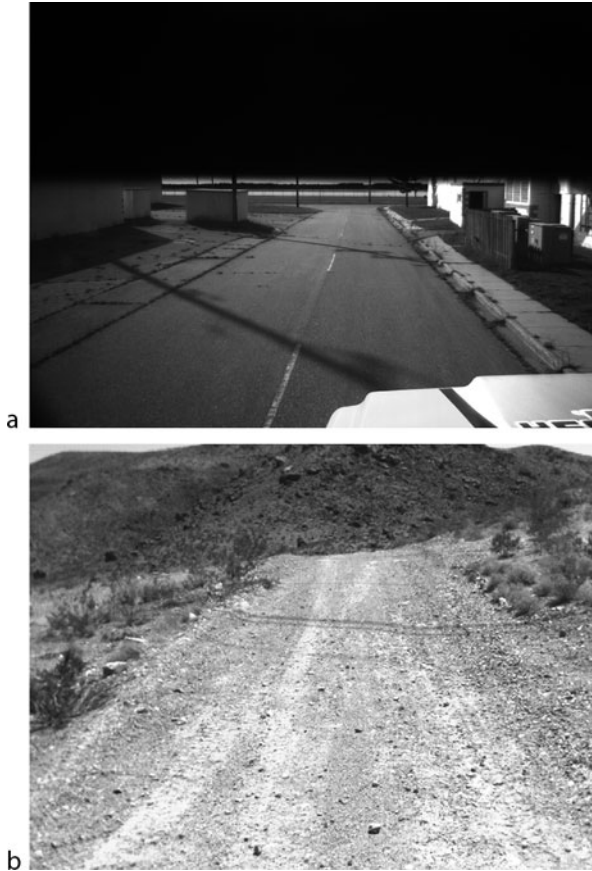
Many different obstacle detection systems have been developed. The main differences between these systems are the types of algorithms and sensors employed. Many studies have focused on road obstacle detection in order to perform such tasks as pre-crash, collision mitigation, stop and go, obstacle avoidance, and inter-distance management. An important issue in ensuring the reliability of obstacle detection is the choice of sensors: digital cameras, infrared sensors, laser scanners, radar, and sonar are commonly used to provide a complete representation of the vehicle's surrounding area, allowing interaction with the world. Inertial sensors like odometers, speed sensors, position sensors, accelerometers, and tilt sensors are used to monitor the motion of the vehicle, measuring its speed, orientation, and position. This chapter is structured in three main sections: The first will introduce a classification of all perception sensors that can be used in this field; a brief description of each sensor will be provided in order to underline pros and cons regarding the obstacle detection field. In the second section, the main algorithms of obstacle detection will be shown, classified by the kind of sensor (or sensors) employed. The third section presents obstacle detection systems that use sensory fusion combining artificial vision with distance detection sensors like laser or radar.

1 Introduction

For autonomous navigation of vehicles in any kind of environment, a map with all the obstacles in the area is necessary in order to allow the vehicle to choose the best trajectory to take. Obstacle detection is a very complex process because of the number of problems that can occur. There are big differences between obstacle detection in a road environment and in other kinds of environments like off-road, as shown in  Fig. 41.1.

In road environments every surface higher than the floor can be considered as obstacle. This definition, that comes from the *flat ground* hypothesis (Talukder et al. 2002), makes the detection of all the elements in the scene much easier. In urban environments the large number of possible obstacles in the scene, especially, when computers trying to recognize their behavior, is the biggest issue. Pedestrians', animals', or other vehicles' behavior is difficult to foresee because of the unpredictable logic behind their movements.

In an off-road environment there are fewer problems because there are fewer possible obstacles in the scene and they are mostly static. The flat-ground hypothesis is not valid in this environment because the profile of the floor can be very complex, requiring a deeper analysis than in the urban environment. Moreover, to distinguish between real obstacles and the elements that identify the limits of the road becomes more difficult because of the problems when trying to identify the correct route without being able to exploit the presence of road lines or asphalt. Another possible issue of the off-road environment is the presence of powder on the floor that may affect data coming from a sensor.



■ Fig. 41.1

A typical urban scene (DARPA Urban Challenge, Autonomous Vehicles in City Traffic [2009](#)) (a) and a typical off-road scene (DARPA's Grand Challenge (Broggi et al. [2006](#))) (b)

Obstacle detection is a difficult process due to the obstacle features' complexity: different shape, color, size, and position, and no one of these characteristics is a priori predictable. Furthermore, in the automotive field obstacle detection algorithms must take in account the prevalence of outdoor scenarios. In this case, there are many environmental issues that affect different sensors: brightness variations do influence vision based systems but do not influence active sensors like laser-scanners or radars. On the other hand, rain or powder may negatively affect active sensors but can be easily overcome on passive sensors like cameras.

Another obstacle detection characteristic in the automotive field is that the observer is moving. This introduces vibration problems that may be strong in an off-road environment, fast background changing, and brightness variations, especially in tunnels or underpasses cross. Moreover, it makes it difficult to recognize the behavior of moving obstacles

surrounding the observer. Algorithms must be designed to provide results fast enough according to the observer speed. An obstacle detection system must (Broggi et al. 2005):

- *Execute in real time*: the data produced by the obstacle detection system should be used by a path-planner; in this case, detection results must be accurate and process fast enough to allow the system time to react;
- *Reliably detect objects*: reliable detection means stable, repeatable, and capable of correctly estimating not only the presence but also the position of obstacles in the scene;
- *Avoid false detection*

2 Sensors for Obstacle Detection

Two parameters can be used to distinguish sensors used for obstacle detection:

- What do they detect?
- How do they detect it?

The first parameter distinguishes sensors as proprioceptive or exteroceptive, while the second distinguishes them as active or passive. This section discusses these two classifications, noting their influence when applied in obstacle detection.

2.1 Proprioceptive and Exteroceptive Sensors

Proprioceptive sensors detect information related to the status of the vehicle, like speed, acceleration, and change of inclination. These data are not directly used while detecting obstacles, but they may become very useful to improve and sharpen the results of the applied algorithms. For instance, to comprehend the behavior of obstacles on the map it is necessary to know the speed; to distinguish small objects and the floor, it is mandatory to know the pitch of the vehicle in comparison with the initial calibration. Proprioceptive sensors are accelerometers, tilt sensors, position sensors, odometers, and, in general, speed sensors. Sensors that collect information about the environment around the vehicle are known as exteroceptive. These sensors collect, directly or indirectly, data about the environment and the obstacles that compose it, their size, their shape, and their distance from the vehicle. Exteroceptive sensors detect the status of the world and allow interactions between the vehicle and the surrounding environment. Exteroceptive sensors include cameras, laser-scanners, and radars.

2.2 Active and Passive Sensors

Sensors are classified as active or passive depending on how they acquire data; in particular, to illuminate and execute detection on an environment, active sensors use

their own energy while passive sensors get the energy from the environment itself. Detection by passive sensors is possible only when there is enough natural energy in the surrounding environment, coming from the sun or from other sources; passive sensors may operate on different wavelengths that differentiate their scope. Among passive sensors, digital cameras are strong candidates for obstacle detection; based on CCD or CMOS sensors, they operate in the visibility spectrum providing images with many more details than those detected using radar or lasers and making obstacle detection easier. Because cameras are passive sensors, they do not emit any signals so there are not interference problems with the environment. Another advantage of passive sensors over active ones is their lower cost, which is why utilization of cameras has increased.

However, cameras are sensitive to lighting conditions: darkness, backlighting, shadows, and poor visibility influence the quality of the images acquired, making the obstacle detection challenging. Moreover, the variability of the scenario could limit detection performance: cluttered backgrounds and unpredictable iterations between traffic participants are difficult to control. The analysis of infrared spectrum for obstacle detection is a widely investigated choice because all heated objects emit infrared radiation that can be registered. According to the operating spectrum range, infrared sensors are divided into near (NIR), middle (MIR), and far (FIR) infrared sensors. An infrared camera provides a thermal image of the scene; since some classes of objects have a specific temperature, object classification can be based on the measurement of the received energy. Different light conditions do not affect the quality of acquired images; the same performance is maintained during the day or night. The main limitation for infrared cameras is that they are sensitive to weather conditions. Heat, rain, and fog influence the heating emissions of objects and make data acquisition unreliable.

Active sensors use their own energy to illuminate the scene, and this very energy is reflected by the obstacles present. The reflected energy becomes the source of information about the status of the environment around the vehicle. The main benefit of active sensors, like laser-scanners or radar, is the possibility of returning to the user a direct measurement of the size and the distance of any object present in the scene simply by measuring the travel time of a signal emitted by the sensor and reflected by the object. The radar technology in obstacle detection systems allows operation at long ranges, in different environmental conditions (rain, snow, fog, low visibility) without many limitations. Moreover, images can be acquired either during the day or at night. However, in complex scenarios like a high-traffic urban road, the radar has limited functionalities because its reliability depends on the radar cross section of the object to be identified. Because metal surfaces are good radar reflectors the vehicles have a much larger radar cross section (10 m^2) than people ($0.2\text{--}2\text{ m}^2$), thus the vehicle detection based on radar technology is easier than human detection.

Laser-scanners are frequently used for obstacle detection, providing high-resolution performance. They operate by sweeping a laser across a scene, measuring the range and returned intensity for each angle. Since lasers work in infrared frequencies the maximum distance for obstacle detection is sensitive to sunlight. Moreover, laser reliability decreases in some weather conditions. In the presence of rain, multiple echoes can lead to worse detection performance. Fog conditions can reflect the laser beam, generating echoes.

Despite the accuracy provided by laser-scanners in automotive systems, a complete representation of the scene is not guaranteed. When a vehicle is pitching up and down the laser beam hits the ground or points to the sky, making data acquisition worthless.

3 Obstacle Detection Methods

A variety of methods for obstacle detection have been proposed: mono vision, stereo vision, active sensor use, and multi-sensor fusion are the principal techniques studied.

3.1 Active Sensor Use

Active sensors allow both obstacle detection and classification: radar performance is not limited to the measurement of distances but other parameters, such as radar cross section, power spectral density and object velocity, are registered and compared in order to discriminate among classes (Gavrila et al. 2001). With the use of laser-scanners, classification is performed grouping close data points in order to obtain different clusters to classify them according to their characteristics. The use of active sensors provides some measurements directly, limiting computational costs. The main drawbacks concern low spatial resolution, slow scanning speed, and interference problems.

3.2 Vision

Mono vision is usually employed when exploiting symmetries of objects or when there is previous knowledge about obstacle location. The detection is performed by processing the resulting images with shape-based approaches in the visible (Nanda et al. 2003) and infrared domain (Bertozzi et al. 2003; Broggi et al. 2004b). Motion-based approaches are also used in the visible and infrared spectrum for monocular obstacle detection (Binelli et al. 2005; Broggi et al. 2004a). The processing of a single image involves low computational costs, but the system performance is limited. The extraction of information about objects' position from a single image is not accurate; thus mono vision is not suitable for obstacle avoidance. Stereo vision is a common choice for obstacle detection systems because it provides a complete three-dimensional representation of the scene. Typically, gray level images are processed, though a few authors dealt with colors. Stereo vision is computationally costly and a compromise between detection range and accuracy must be made. Moreover, the robustness and accuracy of stereo-based methods depends on the parameters obtained from cameras: vehicle motion and windy conditions cause vibrations that affect the estimation of stereo parameters. Generally, two basic steps are computed in a vision-based obstacle detection system: detection and classification. Detection concerns the localization of all the regions containing potential objects and the verification of their correctness; the classification allows determination of the type of

detected candidates. A key problem is the choice of the segmentation method used to extract obstacle features from the background. In a stereo system, segmentation allows determination of which objects of the left image appear in the right one, in order to find homologous points (i.e., the projection of the same world point in the two images) used to compute the disparity. In a mono vision system, segmentation is used to determine a set of features to track. Information about symmetry, color, shadows, and edges can be used for obstacle segmentation. The evaluation of symmetries for obstacle detection involves computation of intensity or edge maps. An important issue to consider with the use of intensity information is the presence of homogeneous areas. They introduce noise that affects the symmetry estimation. Moreover, both intensity and edges informations are sensitive to the variability of the scenario: symmetrical background objects, partially occluded elements, and shadows can lead to false detections. Chromatic information is used for segmentation when the obstacles to detect have distinctive colors. Because color is sensitive to illumination condition, during different moments of the day or under different weather conditions the color of an object can appear different, making segmentation complicated. For this reason, in outdoor settings color segmentation is not a suitable approach. Illumination and weather conditions must also be considered when obstacle segmentation is based on shadow processing. Because the intensity of a shadow is sensitive to the light, there is no systematic way to choose appropriate threshold values. Horizontal and vertical edges are generally detected by searching for changes of intensity in gray or color images. Their use in obstacle segmentation can improve performance. However, the choice of various parameters influences system robustness. Some sets of values, such as the thresholds used for edges detectors or the threshold used to pick the most important edges, may work perfectly under some conditions but might fail in others. Generally, during image the segmentation process the generation of false negatives and false positives involves two main problems that affect system robustness: under-segmentation and over-segmentation. In the case of under-segmentation, some features are not correctly recognized, causing the presence of false negatives, and separated regions are wrongly clustered in a unique region. If over-segmentation occurs, false negatives will be generated (i.e., elements wrongly detected as features) and a single region will wrongly be divided into several. Stereo segmentation approaches are classified in three categories: area-based, featured-based, and pixel-based methods. Area-based algorithms compare sets of pixels within a neighboring window in order to find the best correlation for the disparity computing. The main problem is determining the optimal size for the correlation window. Some obstacle detection approaches use featured-based methods to segment images. Road and vehicle features such as lane markings, shadows, bumpers, windows, and edges are considered in order to find matches for the disparity calculation. The principal drawback of this approach is that it can lead to the generation of sparse disparity maps. When edges of objects are partially visible or missing the feature matching is not performed, which creates holes in the disparity map. Other systems use pixel-based methods for image segmentation. They formulate the disparity problem as an energy-minimizing problem and mathematical functions, like sum of squared difference, are applied for stereo matching.

3.3 Multi-sensor Fusion

Multi-sensor fusion for obstacle detection involves the employment of different sensors together in order to use the solutions provided by a technology when others fail. Because there is not a perfect sensor that provides high performance in any weather and illumination condition, the idea is to use the advantage of one sensor to suppress the disadvantages of another one. In general term, the main targets of sensor fusion are:

- Increasing sensor accuracy in a specific area;
- Extending sensing coverage;
- Increasing result reliability;
- Obtaining more information from sensor's correlation;
- Getting an equivalent, or even more robust, sensor from the fusion of several cheap elements instead of using a single more expensive one.

Obviously, not all of these targets are always reached with fusion, but it is important to understand the power of this method. To perform sensor fusion, it is also important to take into account how sensors are work together, i.e., the sensor network. Three of the most important kinds of sensor networks are:

- *Redundant*: two or more sensors (usually of the same type) get data from the same area. This is useful in systems where it is not possible to make a second measurement on the same scene (for example, with a moving vehicle) or when a backup is required. The network can function when there is at least one working sensor.
- *Complementary*: two or more sensors cover different areas (not only meant as a portion of space, but as sensing capabilities as well), and through fusion it becomes possible to have a wider environmental description. This is the case with different kinds of measurements of the same object at the same time (for example, position and speed).
- *Cooperative*: it is possible to get additional information from fusion, in a way that cannot be acquired from one sensor only. For example, computing the position of one object starting from three distance measurements obtained from different sensors.

Depending on the type of combined information, sensor fusion can be at a low or a high level. At a low level, the fusion involves data information and the results are combined before making a decision. With high level fusion, a list of objects is processed and the fusion is made after some decisions about obstacle size and position are taken. Usually sensor fusion for automotive purposes, for both road safety and autonomous vehicle projects, is performed at a high level. For example, (Labayrade et al. 2005) presents a collision mitigation system, based on laser-scanner and stereo-vision. Data fusion in this case is executed after having extracted obstacles from both sensors and raw data. The information is used to evaluate the time-to-collision for each detected obstacle. The fusion involves the combining of different modules. Results from one sensor are verified using the results of another one. Combining different technologies provides better performance. Since the search space is reduced, the evaluation speed increases, leading to lower computational costs. Because active sensors provide accuracy and robustness in

different weather conditions but limited spatial informations in comparison with passive sensors, a fusion between active and passive sensors leads to better recognition performance. For example, radar and vision sensors are used for a fusion application in dense city traffic situations. Sensors fusion using infrared cameras and radar allows support of the driver in case of reduced visibility, at night, or in adverse weather and illumination conditions. The main advantage of using infrared cameras instead of visible spectrum cameras is that the segmentation problem is simplified.

4 Conclusion

A good obstacle detection system must be capable of performing detection with acceptable computational costs and, at the same time, provide high accuracy, avoiding the presence of false positives or false negatives. The implementation of a reliable obstacle detector is a challenging issue. Any technologies have limitations, especially in different lighting and weather conditions. Moreover, the huge variability of scenarios makes detection a very difficult task. Illumination variations, complex outdoor environments, unpredictable interactions between traffic participants, and cluttered backgrounds are difficult to control.

References

- Bertozzi M, Broggi A, Graf T, Grisleri L, Meinecke M-M (2003) Pedestrian detection in infrared images. In: *Proceedings of the IEEE intelligent vehicles symposium 2003, Columbus*, pp 662–667
- Binelli E, Broggi A, Fascioli A, Ghidoni S, Grisleri P, Graf T, Meinecke M-M (2005) A modular tracking system for far infrared pedestrian recognition. In: *Proceedings of the IEEE intelligent vehicles symposium 2005, Las Vegas*, pp 758–763
- Broggi A, Bertozzi M, Chapuis R, Fascioli FCA, Tibaldi A (2004) Pedestrian localization and tracking system with Kalman filtering. In: *Proceedings of the IEEE intelligent vehicles symposium 2004, Parma*, pp 584–589
- Broggi A, Fascioli A, Carletti M, Graf T, Meinecke M-M (2004) A Multi-resolution approach for infrared vision-based pedestrian detection. In: *Proceedings of the IEEE intelligent vehicles symposium 2004, Parma*, pp 7–12
- Broggi A, Caraffi C, Porta PP, Zani P (2006) The single frame stereo vision system for reliable obstacle detection used during the 2005 Darpa grand challenge on TerraMax. In: *Proceedings of the IEEE international conference on intelligent transportation systems 2006, Toronto*, pp 745–752
- DARPA Urban Challenge, Autonomous Vehicles in City Traffic (2009) Springer Tracts in Advanced Robotics. Springer-Verlag, Berlin Heidelberg
- Gavrila D, Kunert M, Lages U (2001) A multi-sensor approach for the protection of vulnerable traffic participants the PROTECTOR project. In: *Instrumentation and measurement technology conference, 2001. IMTC 2001. Proceedings of the 18th IEEE, vol 3. Budapest*, pp 2044–2048
- Labayrade R, Royere C, Aubert D (2005) A collision mitigation system using laser scanner and stereovision fusion and its assessment. In: *Intelligent vehicles symposium, 2005. Proceedings of IEEE, Las Vegas*, pp 441–446
- Nanda H, Benabdelkedar C, Davis L (2003) Modeling pedestrian shapes for outlier detection: a neural net based approach. In: *Proceedings of IEEE intelligent vehicles symposium 2003, Columbus*, pp 428–433
- Talukder A, Manduchi R, Rankin A, Matthies L (2002) Fast and reliable obstacle detection and segmentation for cross-country navigation. In: *Intelligent vehicle symposium, 2002. IEEE, vol 2. Versailles*, pp 610–618

42 Perception Tasks: Traffic Sign Recognition

Pier Paolo Porta

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>The System</i>	1044
2	<i>Color Analysis</i>	1045
2.1	Color Segmentation	1045
2.2	Chromatic Equalization	1047
3	<i>Shape Detection</i>	1049
3.1	Bounding Boxes Merge and Split	1049
3.2	Pattern Matching	1050
4	<i>Shape Detection Based on Sobel Phase Analysis</i>	1050
4.1	Edges Detection	1051
4.2	Analysis of the Sobel Phase Distribution	1051
5	<i>Classification</i>	1053
5.1	Neural Network	1054
5.2	Tracking	1056
6	<i>Output and Results</i>	1056
7	<i>Conclusions</i>	1059

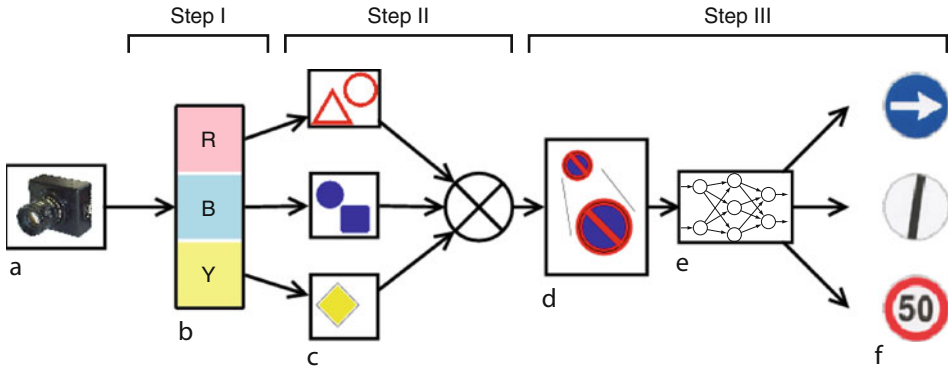
Abstract: The system described in this chapter is a traffic sign recognition based on a color camera. Each algorithm step will be detailed: a color segmentation to identify the possible regions of interest, a shape detection, and the final sign classification and tracking. A description of the encountered problems and their solutions is given as well. The last section presents the algorithm results.

1 The System

Automatic traffic signs detection and classification is a very important issue for Advanced Driver Assistance Systems (ADAS) and road safety: different road signs detectors were developed in the last 10 years (Nguwi and Kouzani 2006). Most of the industrial systems developed are based only on speed limit sign recognition, on the contrary the system proposed here can detect a large scope of road signs. The system described here (Broggi et al. 2007) works with a camera already mounted on-board for other purposes such as lane departure warning (LDW). Another advanced feature introduced here is the low dependence from illumination conditions: this is of paramount importance for good performance in early mornings and late afternoons where sunlight usually presents an appreciable deviation toward red.

Missed signs can cause dangerous situations or even accidents. An automatic road sign detection system can be used both to warn drivers in these situations and to supply additional environmental information to other onboard systems such as ACC (Adaptive Cruise Control).

Both gray-scale and color cameras can be used for this purpose; in the first case, the search is mainly based on shape and can be quite expensive in terms of computational time (Gavrila 1999; Loy and Barnes 2004). Using a color camera, the search can be based mainly on color: color segmentation is faster than shape detection, although requiring additional filtering. Images acquired by an inexpensive color camera can suffer from Bayer conversion artifacts and other problems such as color balance, but anyway, the developed system is more robust and definitely faster. Some research groups have already used color images for traffic signs detection. Most of these methods have been developed using color-based transformations; HSV/HSI color space is the most used (de la Escalera et al. 2003; Vitabile et al. 2001) but other color spaces, such as CIECAM97 (Gao et al. 2002), can be used as well. These spaces are used because chromatic information can be easily separated from the lighting information: this is used to detect a specified color in almost all light conditions. Anyway traffic signs can be detected in RGB (Soetedjo and Yamada 2005) or YUV (Shadeed et al. 2003) color space with the advantage that no transformation, or just a very simple one, is required. The segmentation and thresholding algorithms are anyway more complex, but a lot of computational time can be saved. In order to make the detection more robust, both color segmentation and shape recognition can be used in cooperation (Gao et al. 2002). Anyway the processing described so far has to be computationally light to keep the advantages of the selected color space.



■ Fig. 42.1

Algorithm flowchart. (a) Image acquisition; (b) color segmentation; (c) shape detection; (d) bounding box scaling; (e) classification; (f) output

Many different approaches are used for the subsequent classification: most of them are based on artificial intelligence techniques; the most used are neural networks (de la Escalera et al. 1994; Hoessler et al. 2007), bayesian networks, and fuzzy logic (Jiang and Choi 1998).

The proposed approach (Broggi et al. 2007) is based on three steps (see ► Fig. 42.1): color segmentation, that is presented in the following section, shape detection (● Sect. 3), and classification (● Sect. 5) based on several neural networks.

2 Color Analysis

In this section, the first step of road signs detection based on color information is presented. One of the main road sign features is their immediate identification by a human driver: this is due to a limited number and very specific set of shapes and colors. In particular, the color set used for road signs is composed of white, gray, red, yellow, and blue; green used on highway signs (in Italy) is not considered in this analysis.

This section presents a discussion on the choice of the color space to use, a robust color segmentation, and a solution for the problem of chromatic predominance of the light source.

2.1 Color Segmentation

The specific illumination condition deeply affects the road sign color perception. Common environmental conditions are usually characterized by a wide range of different illuminations: direct sunlight, reflected sunlight, shadows, and sometimes even different illuminations can coexist on the surface of the same sign as shown in ► Fig. 42.2.

The objective of this work is to identify a sign of a given color (for example red) regardless of its illumination. As already mentioned, in literature most approaches are



■ Fig. 42.2

Examples of different illumination conditions

based on HSV or HLS color space (de la Escalera et al. 2003; Jiang and Choi 1998; Shadeed et al. 2003) but the camera used for this application has a raw Bayer output and a conversion would be too computationally expensive because of the nonlinearity introduced. Therefore, RGB space is too dependent on brightness, so different color bases were tested to find one that is less dependent on brightness. All the color base conversions evaluated can be obtained with linear transformations; this kind of approach has been followed also by de la Escalera et al. (1994, 1997), Jeun-Hai and Gang (1994), Shadeed et al. (2003), and Soetedjo and Yamada (2005).

First YUV color space has been tested: the Y coordinate is strongly dependent on brightness thus, considering only the UV plane, it is possible to bound regions mainly based on hue.

Empiric tests demonstrated that this kind of bounding is too simple to collect all the cases of different illumination and to cover most of the case study.

The second attempt focused on RGB values: what is bounded here is the ratio between different channels and not only the channel itself. Equation 42.1 shows the expression of this kind of thresholding:

$$\begin{cases} \alpha_{min} * G < R < \alpha_{max} * G \\ \beta_{min} * B < R < \beta_{max} * B \\ \gamma_{min} * B < G < \gamma_{max} * B \end{cases} \quad (42.1)$$

The values of the parameters involved have been obtained tuning the algorithm on real images in different conditions.

Both these methods have produced good results but the second one is chosen for several reasons:

- Lower number of false positives.
- Lower computational load: indeed no color space conversion only a thresholding is needed.
- Easier to tune: because it is less sensible to small parameters variations.

After an analysis on colors present in Italian traffic signs, the research has been focused on three colors: red, blue, and yellow; the described algorithm is applied on all these three colors to generate three binary images containing only the pixels referred to that color.

All the pixels referring to a same connected region are labeled together. Regions smaller than a fixed threshold are discarded because they usually do not represent a road sign or, otherwise, the target is so far that a detection would not be followed by a successful classification due to reduced size.

Considering, for example, the red color segmentation, all the red objects present will be detected (e.g., cars, buildings, placards), but these detections will be discarded by the following further steps.

2.2 Chromatic Equalization

One of the main problems experienced in this stage is the dependence on the color of the light source. For example, during sunset or dawn a red color predominance is present and this deeply affects the color segmentation step; see [Fig. 42.3](#).

To solve this problem a chromatic equalization based on two steps has been developed:

- Light source color identification
- Chromatic correction

The easiest way to find the light source color is to find an object supposed to be white and then compute the aberration from theoretical white (255, 255, 255 in the RGB color space). Unfortunately on a dynamic environment such as roads, it is difficult to have a white reference point. Thus, the color of the road is searched for, as suggested in Buluswar (2002) that is supposed to be gray. In Buluswar (2002) the chromatic response



■ Fig. 42.3

Original image and result of chromatic equalization. In the *right* image the region supposed to frame the asphalt is shown

of several materials has been analyzed in different illumination conditions. Moreover, in most cases, in vehicular application a specific region of the image frames the asphalt as shown in [Fig. 42.3](#).

Through the use of a temporal window in which the light source color is integrated, it is possible to avoid fast changing of the result and keep it stable. In case where tracking has to be introduced in the processing chain, it is very important to have stable conditions for a reasonable number of frames.

Once the light source color has been evaluated, chromatic equalization can be applied. This step is very similar to a gamma-correction process: to reduce the computational time a linearization of the gamma-correction function has been used, as shown in [Fig. 42.4](#) and described below.

Line A: $y = \alpha \cdot x$

Line B: $y = \beta \cdot x + \gamma$

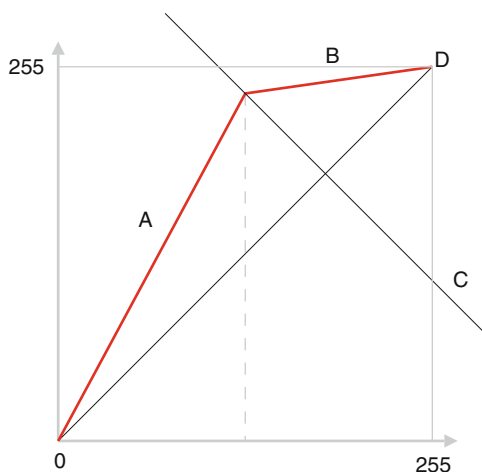
Line C: $y = -x + k$

D: point of coordinates (255, 255).

Now suitable values for the parameters α , β , γ , and k have to be found. The α value for the three channels can be computed in three steps:

- Consider the RGB value of the light source color
- Set $\alpha = 1$ to the channel that has the intermediate value
- Set $\alpha = \frac{\text{intermediate channel value}}{\text{channel considered value}}$ for the other two channels.

The parameter k can be set once with empiric tests, to avoid saturations, β can be obtained forcing the curve to be continued and to reach D.



■ Fig. 42.4

Gamma-correction curve

3 Shape Detection

After color segmentation, a first sorting based on shape is performed. This sorting is developed in order to reduce the complexity of the final classification. Two different methods are used to determine the correct shape with high reliability; the first one is based on pattern matching, the other one is based on remarks about edges. Before this sorting, a method to merge and split the bounding boxes generated by color segmentation is applied together with a filtering based on aspect ratio.

Another method for shape detection based on Sobel phase analysis will be described as well. This approach is more robust to sign rotation but is not suitable for circle shape detection.

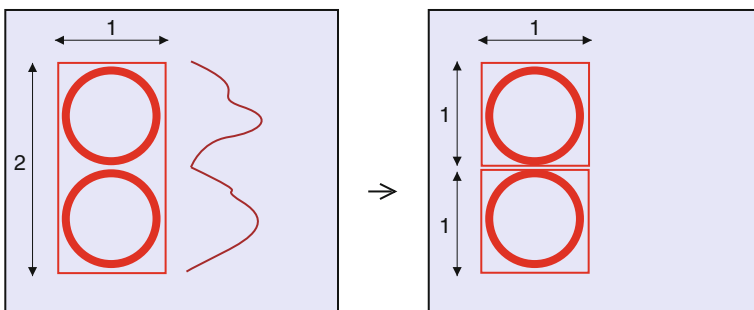
3.1 Bounding Boxes Merge and Split

Color segmentation can sometimes provide a bounding box that contains two or more signs or only a part of a sign; the use of such a bounding box in the subsequent classification may cause an error: this merge and split step is developed in order to solve this problem.

When two signs with the same color are hunged on a single pole it may happen that the segmentation identifies the two signs as a single one because of the weak separation between the signs. All the boxes with height almost double as width are checked considering the vertical histogram of the binarized image. If a very low value is identified around the middle of the histogram, the bounding box is splitted, see [Fig. 42.5](#).

On the other hand, bounding boxes of the same color, that overlap more than a threshold, are merged together as a single sign. This process is useful to merge in a single box, different parts of a same sign that may have been divided by the labeling step.

A single sign can also contain two colors that can be identified in the two corresponding images, for example, work in progress signs contain both red and yellow.



■ Fig. 42.5
Bounding boxes split

All the three color segmentation images are checked for overlapping bounding boxes and, if overlapping bounding boxes are found, they are merged into a single bounding box. Each bounding box can handle up to two colors, a primary and a secondary one; if a bounding box is the result of a merging, its primary color will be set to the color of the larger bounding box while the other bounding box will be set to the secondary color. If a bounding box does not overlap any other box, it will have only a primary color.

3.2 Pattern Matching

Bounding boxes with primary color red or yellow are supposed to have specific shapes that can be easily detected using a pattern matching. A reference pattern is built for each shape that has to be detected. The patterns are built based on the shape of signs, with the aim of detecting rotated or misaligned signs also. Triangle, reversed triangle, circle, and filled circle are searched in red bounding boxes, while only rhombus is searched in yellow ones. Filled circles are used to detect stops and no thoroughfare signs.

All the detected bounding boxes are resampled to a fixed size (50×50 pixels) equal to the pattern size. A very simple pattern matching is used in order to reduce the complexity and the computational time.

4 Shape Detection Based on Sobel Phase Analysis

Another method for shape detection has been investigated, based on the evaluation of the Sobel edges and Hough images in a region of interest detected by the color-based stage. The system proposed is a simple shape detector independent from geometric distortion, that is, rotation, partial occlusion, deformation, and translation.

Many approaches are presented in literature for shape detection: Soetedjo and Yamada (2005) and Zheng et al. (1994) show that pattern matching is a robust and fast method; in Barnes and Zelinsky (2004) Hough transformation based on radial symmetry is used to detect speed signs. Genetic algorithms, used in Aoyagi and Asakura (1996) and de la Escalera et al. (2001, 2003, 2004), allow accurate results in shapes detection, but their execution requires computational times unsuitable for real-time applications. Other methods (Gil-Jimenez et al. 2007; Lafuente-Arroyo et al. 2005; Maldonado-Bascon et al. 2007) are based on the use of supervised learning methods for classification, like SVM. Some methods recognize only a specific shape: de la Escalera et al. (1994, 1997) and Gao et al. (2002) present a triangular shape detector; in Barnes and Zelinsky (2004) and Soetedjo and Yamada (2005) only circular shapes are detected; the algorithm presented in Loy and Barnes (2004) allows triangles, rectangles, and octagons detection, through the use of a fast radial symmetry and the shape-center detection. Broggi et al. (2007) have been presented a vertical traffic sign recognition system based on a three-step algorithm: color segmentation, shape recognition, and neural network classification to detect and classify almost all Italian traffic signs. The shape detection method described was sensible to

translation and rotation of signs. To address this problem the proposed approach detects edges on images, correlating them with their gradient distribution: the shape of a placed object depends on the edges and on their features that is, position and mutual orientation.

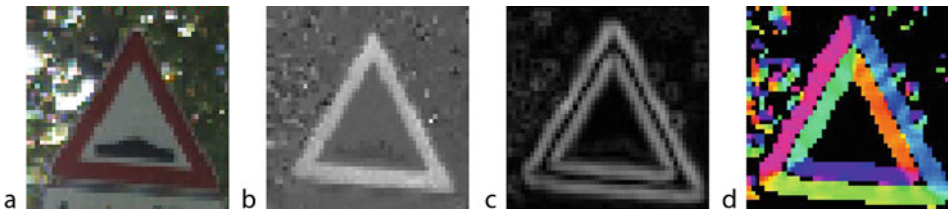
4.1 Edges Detection

A conversion from RGB to gray scale images is performed to detect edges: each region limited by a bounding box is cropped from the color image (► Fig. 42.6a) and converted to a gray scale crop (► Fig. 42.6b). Since the RGB to gray conversion causes information loss due to the passage from a vector field (color image) to a scalar field (gray image), in order to increase the $\frac{\text{signal}}{\text{noise}}$ ratio, the predominant color component, according to the primary color of each bounding box detected, is extracted from the original crop and pointed out in that gray.

In order to enhance edges, Sobel operator with a 5×5 mask is applied for each gray scale crop, obtaining images about the Sobel norm (► Fig. 42.6b) and the Sobel phase (► Fig. 42.6c) intended as $\arctg(G_y/G_x)$.

4.2 Analysis of the Sobel Phase Distribution

Using images obtained by Sobel filtering, analysis about the phase distribution of Sobel edges is performed to detect distinctive features for each shape. This step focuses on the study of the most frequently used edge gradient for each region of interest detected, to decide if the placed object has a road sign compatible shape and which one is. The main purpose of this step is the study of the peaks in the Sobel edges phase distribution, regardless of the sign of the transition described by each edges. Since opposite transitions cause, in the Sobel phases, image angles at a distance equal to π , static results represented in the range $(-\pi, \pi)$ are mapped to the range $(0, \pi)$, because at this step, the study of the edge gradient's most frequent directions is the most interesting phase. ► Figure 42.7



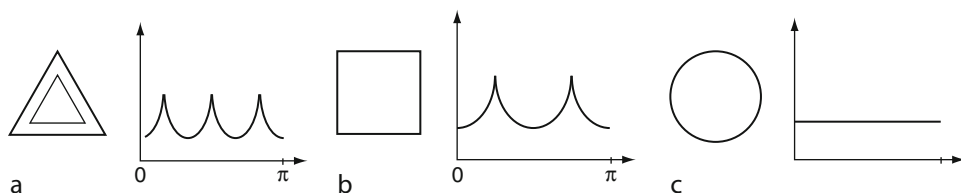
■ Fig. 42.6

Shape detection based on Sobel phase analysis processing steps: (a) Crop of the sign to be detected. (b) *Gray scale* conversion with primary component highlighted (in this case the triangle shape which was *red*). (c) Sobel norm image. (d) Sobel phase image

shows, for each considered shape, the ideal trend of its phase distribution; according to the shape of interest the distinctive features are:

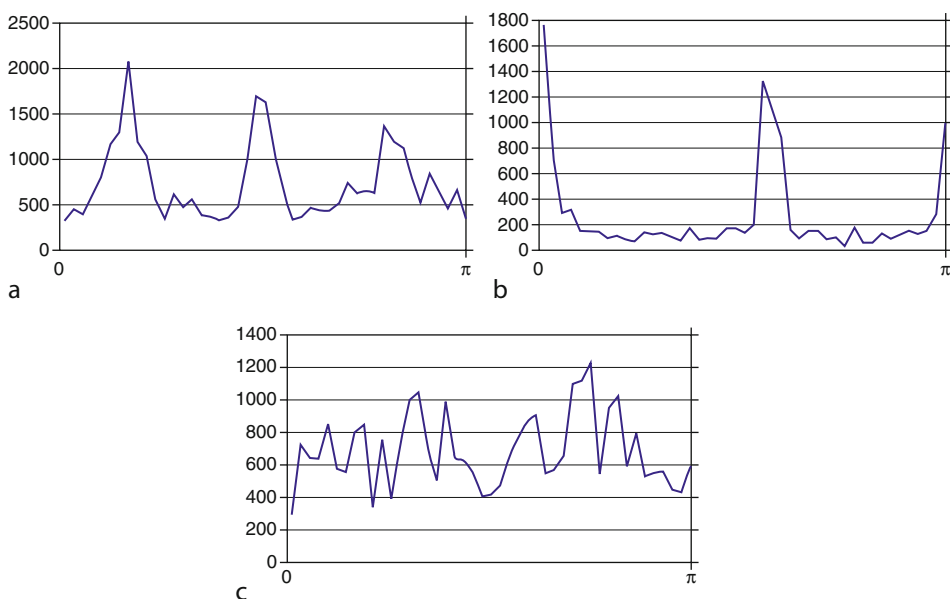
- (a) Three peaks placed at a distance equal to $\frac{\pi}{3}$ for triangles
- (b) Two peaks set at a distance equal to $\frac{\pi}{2}$ for rectangles
- (c) Histogram with all equal values for each considered angle in case of circles

Some examples relative to phase distribution of real signs are reported in [Fig. 42.8](#). It can be noticed that circular sign signature is far from the ideal trend of [Fig. 42.7a](#). This is due to the symbols drawn on the sign and by the elements placed on the background that



■ Fig. 42.7

Histogram of the ideal Sobel phase angles distribution, for the outlines of (a) concentric triangles, (b) rectangles, and (c) circles



■ Fig. 42.8

Histograms of the Sobel angle phase distributions for: (a) triangular signs, (b) rectangular signs, and (c) circular signs. For each considered phase angle (in the range $[0, \pi]$), the number of pixel having that phase or its supplementary is computed

introduce significant oscillations on the phase histogram involving a trend different from the ideal constant one. The same noisy elements affect also the phase distribution of triangular and rectangular shapes, but in these cases the distinctive peaks of the phase histogram are greater than the oscillations introduced by the noisy components.

Statistical analysis on the real signs is performed in order to evaluate, as a distinctive feature, the real gap among peaks in the Sobel phases distribution according to the shape delimited by the bounding box. The distance between two peaks is estimated as the minimum of circular distance, calculated in clockwise and anticlockwise directions:

$$\min(\text{abs}(v1 - v2), \text{range} - (\text{abs}(v1 - v2)))$$

Since the minimum circular distance between two peaks in the range $(0, \pi)$ could vary between 0 and $\frac{\pi}{2}$, for the analysis of the real gap discrete intervals of the range $(0, \frac{\pi}{2})$, that is, the x-axis is split into 23 discrete intervals of width equal to $4\frac{\pi}{180}$ (4°) are considered.

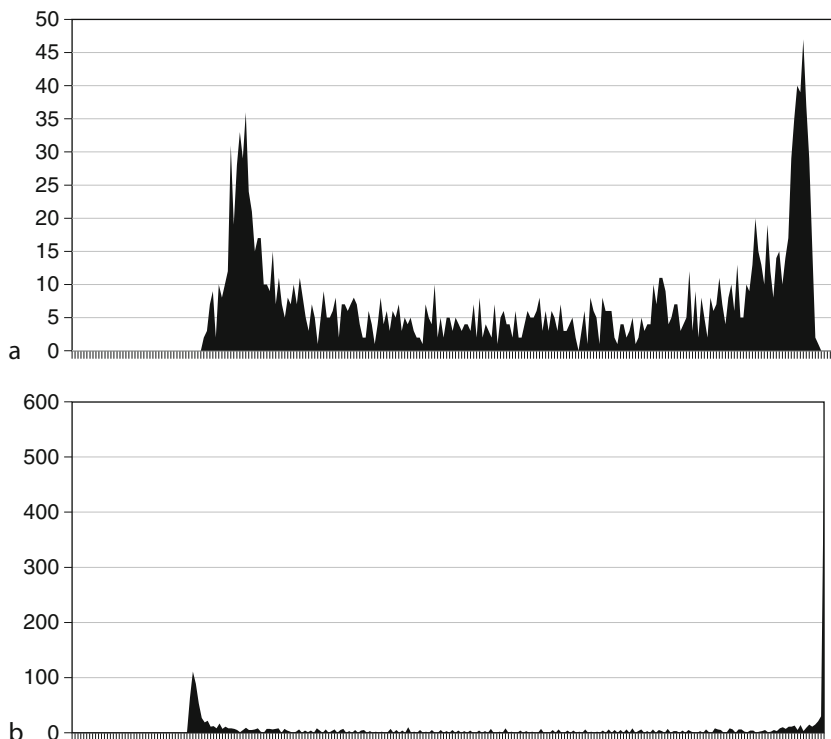
According to the expectations, bounding boxes with triangular shape are concentrated in a neighborhood of the ideal peak gap value, that is, 60° : for the selection of the largest number of regions looking for triangles, all bounding boxes having a peak gap in a range on the left and on the right of ideal peak gap value are considered as possible candidates. The range value has been experimentally computed to be 12° on each side. Bounding boxes with rectangular shape present a peak gap value close to the last gap sample considered equal to $\frac{\pi}{2}$. The threshold to forward a bounding box to the rectangles processing is in the neighborhood equal to the three last represented intervals.

5 Classification

Regions provided by the sign detector are converted to grayscale and resampled to 50×50 pixels. Since on this regions intensity histograms will be analyzed, and the nearest interpolation can generate an imprecise histogram, bilinear interpolation becomes necessary during the resampling process.

In road signs composed by a border and an inner symbol (Danger and Prohibition), only the internal region is resampled to the target size, while the other signs are resampled in their entirety. A mask is then applied on the resampled regions in order to remove the background and the border where present. This mask is generated using shape information (for example, a circular mask for Prohibitions and Obligation) and adapted to the border.

From this filtered region, a 256 bin gray scale histogram is computed. Since a mask has been applied, the number of pixels analyzed is reduced by up to 50%, but it is still large enough to have a reliable histogram without a filtering pass. On this histogram two peaks are selected, one relative to the background and one relative to the shade of the symbol. For road signs with white inner symbol (Obligation), pixel values are inverted, in order to represent all symbols with a black shade and white background, independently of their class. Different sign classes have different ratios between background and inner symbol size; therefore, depending on the specific sign class a different peak detector is used to find the shade indexes for the background and the symbol.



■ Fig. 42.9

Illumination histogram of region provided by detection stage, before (a) and after (b) contrast stretching

The two detected peaks are used to contrast stretch the filtered region between 40 and 255 (see ► Fig. 42.9), with the bottom value (40) chosen according to the black level of synthetic test images. Some examples of refined and illumination corrected regions are shown in ► Fig. 42.10.

The presented algorithm allows to compensate for scale, position, and intensity variations, while noise due to occlusions, dirt, or other environmental conditions are not handled by the presented processing steps.

The final reduced and normalized regions of any traffic sign candidate are directly used as the input vector of a neural network classifier, which is presented in the following section.

5.1 Neural Network

A multilayer perception network with feed forward topology is used for classification, and six different neural networks have been trained, one for each reported class of road signs. No special network output is reserved for false positives, since they are expected to provide

low values on all output nodes. Networks are trained with the back-propagation method, and validation sets are used to avoid overfitting.

Neural network topologies differ among classes. The input layer is always 50×50 large, and output layer size is equal to the number of signs associated with the given category. The number of hidden layers (one or two) and their size have been chosen to optimize the network performance after extensive benchmarks: the use of only a few neurons and a hidden layer usually does not give satisfactory results, while too many of them causes overfitting. Neural networks with different topologies (with different numbers of layers and layer sizes) have been trained and tested. The best network for each class has been adopted for use in the final application. Currently, adopted topologies are shown in [Table 42.1](#).

A back-propagation approach is used to train the networks, using both synthetic patterns (roto-translated synthetic signs) and real ones, manually chosen in order to cover a broad range of cases, but trying to avoid excessive specialization of the network: in order to avoid that kind of specialization only a portion (about 5–10%) of real signs is used to train the networks.

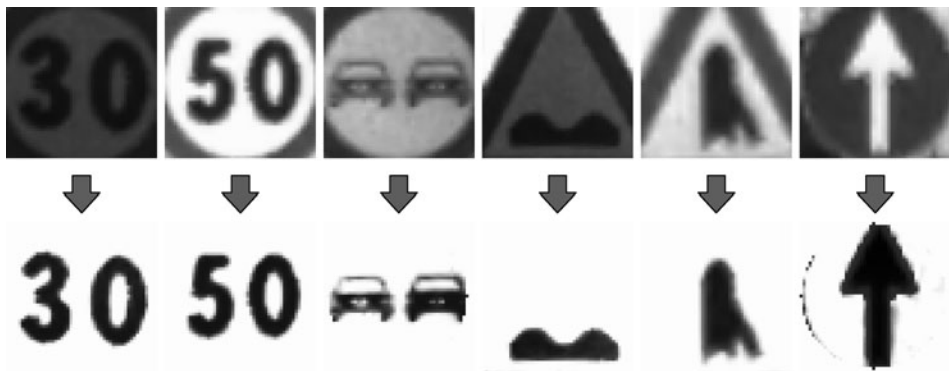


Fig. 42.10
Contrast stretching and filtering applied to some test regions provided by detection stage. Illumination variations are rejected and border of sign partially removed

Table 42.1
Neural network geometry

Net name	Number of Hidden layer	Geometry	Number of output
Prohibitions	2	115 + 65	37
Information	1	50	10
Obligation	2	90 + 40	26
Danger	1	175	42
No-Parking	1	40	3
Stop	1	80	3

The neural networks are functions with 2,500 (50×50) inputs and several outputs; each output represents the confidence of classification of the associated sign. Outputs of the network are further processed during the tracking stage described in the following section.

5.2 Tracking

The use of a tracking stage is important for several reasons:

- Neural networks output is not stable and can occasionally provide wrong results.
- In order to provide a useful road sign classification system, only one output for each encountered sign should be reported.

Before classification all the regions selected by the detection stage are compared with those classified in the previous frames. For this reason only successfully classified regions can be tracked.

In case of a successful match, regions are tracked, while signs remaining without a match are marked as “ghost.”

Neural networks can be considered as transfer functions with several inputs, and n outputs $O_0 \dots O_n$. In any frame, outputs of the neural network are averaged with the previous network output $A_0 \dots A_n$ of the same tracked sign, using as weight the region size w , so that far candidates have less weight compared to the closer ones.

$$\begin{aligned} A'_i &= A_i + O_i * w \\ W' &= W + w \end{aligned} \quad (42.2)$$

When a tracked object goes outside of the screen or becomes “ghost” for more than five frames, the sign identifier is reported using the maximum average value

$$id = \max_i A_i \quad (42.3)$$

only if the normalized output value A_{id}/W is above a given threshold ξ , which in general assumes different values for each sign class. Thresholds have been chosen through the use of receiver operating characteristic (ROC) curves analysis.

In order to reduce the system latency, if a road sign is constantly tracked for at least ten frames, it is classified in advance and directly reported to the user.

6 Output and Results

Tests have been performed in several situations, with different illumination conditions.

◆ **Figure 42.11** shows some examples of different types of signs. The black stripe placed below each frame is divided into two lines: the top line shows all the signs detected in that frame scaled to a 50×50 pixel image, while the bottom line shows all the corresponding models for the signs that have been classified.

All kinds of sign are correctly detected, even in some ambiguous cases such as [Fig. 42.11c](#). Generally, signs are recognized when they are relatively close to the vehicle (e.g., 20 m) and appear not too misaligned with the camera; for example, the perspective deformation of the yield sign in [Fig. 42.11f](#) does not allow the system to detect it, while the same sign in [Fig. 42.11d](#) is correctly detected. Empirical tests demonstrated that signs can be detected up to 30 m ahead. However, this distance can be increased reducing camera focal length but dropping the possibility to detect signs that are close and at the side of the vehicle.

During all the development process, low computational time has been one of the issues to follow: on a Pentium 4 at 3 GHz the algorithm runs faster than 10 Hz.

Another open problem is to understand whether a detected sign refers to the driver or not: it can happen, especially in junctions, that a sign is seen by drivers running on another road. This problem can be solved only if the system can perceive the road and junction structure. It is not unusual that a sign is mounted in a wrong way as, for example, shown in [Fig. 42.11d](#).

Regarding the shape detection based on Sobel phase analysis, the algorithm has been deeply tested on different scenario frames leading to robust results: all categories of triangular ([Fig. 42.12a, b, c, d](#)) and rectangular ([Fig. 42.12e, f](#)) signs are correctly detected. The presence of a slightly illuminated sign ([Fig. 42.12c](#)) does not affect the detection performances.

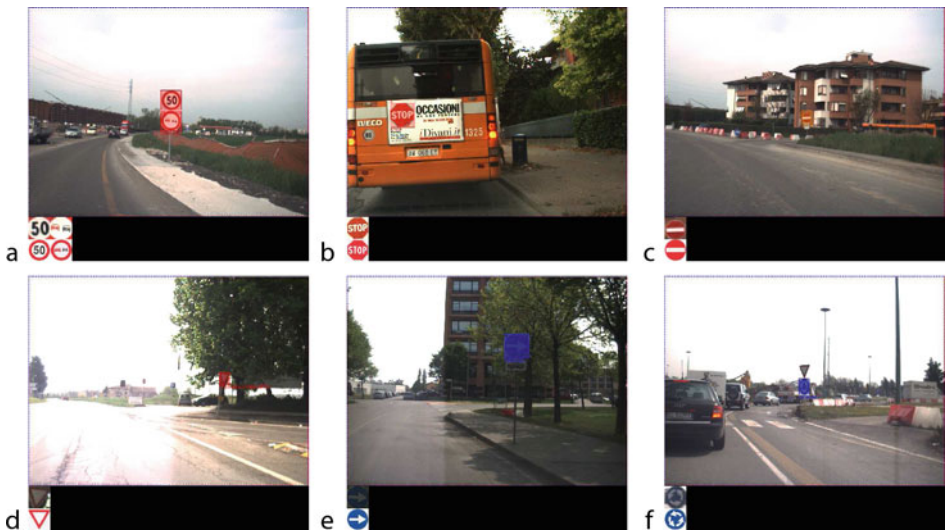


Fig. 42.11

Results on several conditions and different signs. Different bounding box colors indicate primary and secondary color assigned with color segmentation. (a) pool of two signs; (b) the sign with building on the background having the same color; (c) sign on a placard; (d) yield sign in a saturated image and slightly rotated; (e) obligation sign; (f) yield sign is not detected because of its high rotation, the sign below that instead, has been recognized even if it is mounted upside-down

As shown in Fig. 42.13 the presented approach allows the detection of rotated signs, both in the presence of small rotations (Fig. 42.13a, b, c, d) and heavy rotations (Fig. 42.13e, f). The system has been tested in both cases of vertical rotation, due to a misplacement of the road sign, and in the case of horizontal rotation that will result in a rotation due to the perspective effect.

In Fig. 42.14 are shown two cases of false positives where the stripes placed on the trash bin (Fig. 42.14a) and two trees in front of a red building (Fig. 42.14b) generate



Fig. 42.12

Detection of: (a–b) danger signs, (c–d) work in progress signs, (e–f) rectangular information signs

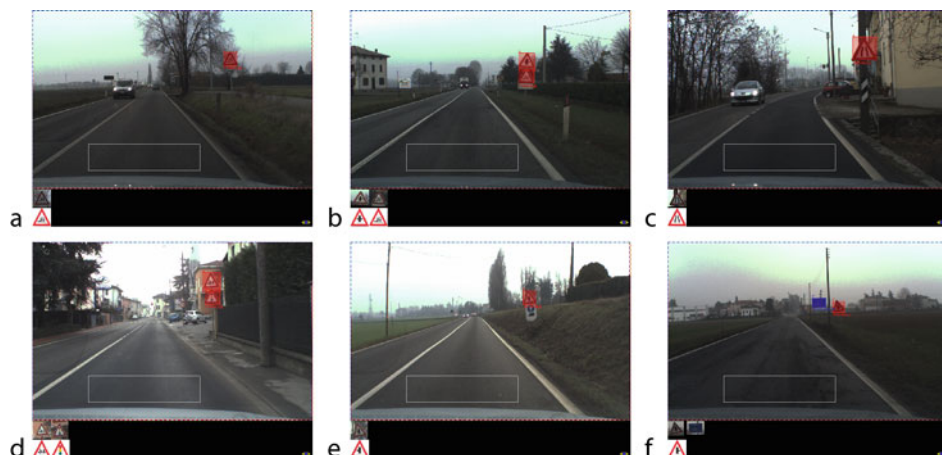
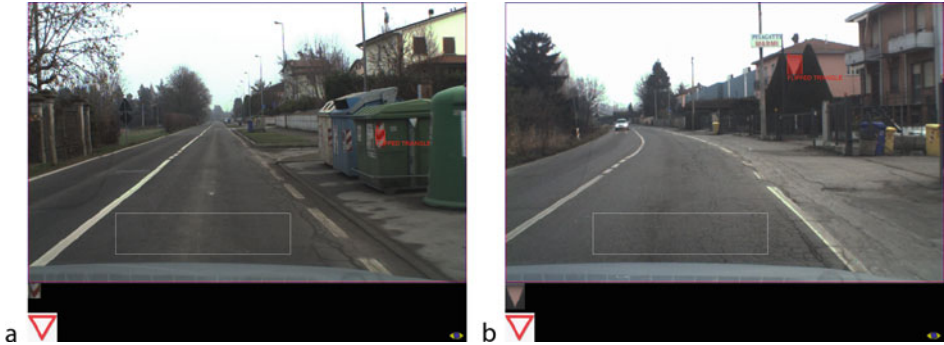


Fig. 42.13

Detection of rotated signs



■ Fig. 42.14
False positives

patterns compatible with the presence of a priority sign. The presence of these false positives could be mitigated analyzing the positions of the elements with respect to the road in order to verify that they are set in a place compatible with that of a road sign.

7 Conclusions

In this chapter, an example of traffic sign recognition system has been described: starting from the preprocessing step to the final classification. Some issues still remain open, such as, for example, the effectiveness of a speed limit, that ends after the first junction, or the ambiguous placement of some signs that lead to possible misunderstandings.

These problems could be mitigated through the use of a GPS with a map of the installed signs. In this scenario the traffic sign recognition system would be highly useful in case of temporary signs: in fact, usually, temporary signs are placed in danger areas, and thus their detection can be really important.

References

- Aoyagi Y, Asakura T (1996) A study on traffic sign recognition in scene image using genetic algorithms and neural networks. In: Proceedings of the IEEE 22nd international conference on industrial electronics, control, and instrumentation, vol 3, Denver, pp 1838–1843, Aug 1996
- Barnes N, Zelinsky A (2004) Real-time radial symmetry for speed sign detection. In: Proceedings of the IEEE intelligent vehicles symposium, 2004, San Diego, pp 566–571
- Broggi A, Cerri P, Medici P, Porta PP, Ghisio G (2007) Real time road signs recognition. In: Proceedings of the IEEE intelligent vehicles symposium, Istanbul, Turkey, pp 981–986, June 2007
- Buluswar SD (2002) Color-based models for outdoor machine vision. Ph.D. dissertation, University of Massachusetts Amherst, 2002
- de la Escalera A, Moreno LE, Puente EA, Salichs MA (1994) Neural traffic sign recognition for autonomous vehicles. In: Proceedings of the IEEE 20th international conference on industrial

- electronics, control and instrumentation, vol 2, Bologna, pp 841–846, Sept 1994
- de la Escalera A, Moreno LE, Salichs MA, Armignol JM (1997) Road traffic sign detection and classification. *IEEE Trans Ind Electron* 44:848–859
- de la Escalera A, Armignol JM, Salichs M (2001) Traffic sign detection for driver support systems. In: 3 rd international conference on field and service robotics, Espoo, Finlandia, June 2001
- de la Escalera A, Armignol JM, Mata M (2003) Traffic sign recognition and analysis for intelligent vehicles. *Image Vis Comput* 21(3):247–258
- de la Escalera A, Armignol J, Pastor J, Rodriguez F (2004) Visual sign information extraction and identification by deformable models for intelligent vehicles. *IEEE Trans Intell Transport Syst* 5(2):57–68
- Gao X, Shevtsova N, Hong K, Batty S, Podladchikova L, Golovan A, Shaposhnikov D, Gusakova V (2002) Vision models based identification of traffic signs. In: Proceedings of European conference on color in graphics image and vision, Poitiers, France, pp 47–51, Apr 2002
- Gavrila D (1999) Traffic sign recognition revisited. In: Mustererkennung 1999, DAGM-symposium. Springer, London, UK, pp 86–93
- Gil-Jimenez P, Gomez-Moreno H, Siegmann P, Lafuente-Arroyo S, Maldonado-Bascon S (2007) Traffic sign shape classification based on support vector machines and the FFT of the signature of blobs. In: IEEE intelligent vehicles symposium, 2007 Istanbul, pp 375–380, 13–15 June 2007
- Hoessler H, Wöhler C, Lindner F, Kreßel U (2007) Classifier training based on synthetically generated samples. In: Proceedings of the 5th international conference on computer vision systems, Bielefeld, Germany, Mar 2007
- Jeun-Haii F, Gang L (1994) A vision-aided vehicle driving system: establishment of a sign finder system. In: Proceedings of the IEEE conference on vehicle navigation and information systems, 1994, Yokohama, pp 33–38, Aug–Sept 1994
- Jiang G-Y, Choi TY (1998) Robust detection of landmarks in color image based on fuzzy set theory. In: Proceedings of the IEEE 4th international conference on signal processing, vol 2, Beijing, China, pp 968–971, Oct 1998
- Lafuente-Arroyo S, Gil-Jimenez P, Maldonado-Bascon R, Lopez-Ferreras F, Maldonado-Bascon S (2005) Traffic sign shape classification evaluation i: Svm using distanceto borders. In: Proceedings of the IEEE international conference on intelligent vehicles symposium, 2005, Las Vegas, pp 557–562, 6–8 June 2005
- Loy G, Barnes N (2004) Fast shape-based road signs detection for a driver assistance system. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, Sendai, Japan, pp 70–75, Sept 2004
- Maldonado-Bascon S, Lafuente-Arroyo S, Gil-Jimenez P, Gomez-Moreno H, Lopez-Ferreras F (2007) Road-sign detection and recognition based on support vector machines. *IEEE Trans Intell Transport Syst* 8(2):264–278
- Nguwi Y, Kouzani A (2006) A study on automatic recognition of road signs. In: Proceedings of the IEEE conference on cybernetics and intelligent systems, Bangkok, Thailand, pp 1–6, June 2006
- Shadeed WG, Abu-Al-Nadi DI, Mismar MJ (2003) Road traffic sign detection in color images. In: Proceedings of the IEEE 10th international conference on electronics, circuits and systems, 2003, vol 2, Sharjah, pp 890–893, Dec 2003
- Soetedjo A, Yamada K (2005) Fast and robust traffic sign detection. In: Proceedings of IEEE international conference on systems, man and cybernetics 2005, vol 2, Istanbul, pp 1341–1346, Oct 2005
- Vitabile S, Pollaccia G, Pilato G (2001) Road signs recognition using a dynamic pixel aggregation technique in the HSV color space. In: proceedings of international conference on image analysis and processing, Palermo, Italy, pp 572–577, Sept 2001
- Zheng Y-J, Ritter W, Janssen R (1994) An adaptive system for traffic sign recognition. In: Proceedings of the intelligent vehicles 1994 symposium, Paris, pp 165–170, Oct 1994

43 Vision-Based ACC

Matteo Panciroli

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Introduction</i>	1062
2	<i>Systems</i>	1063
3	<i>Algorithm Overview</i>	1064
3.1	Vehicle Detection	1066
4	<i>Conclusions</i>	1068

Abstract: In this chapter, a description of ACC system will be given, focusing on how to develop this kind of system using optical sensors and computer vision techniques. In particular, two approaches based on different sensors fusion (LIDAR, radar, and camera) are described.

1 Introduction

The ACC acronym has different origins. It may be considered *Autonomous Cruise Control* or *Adaptive Cruise Control* or even *Active Cruise Control* depending on the different manufacturers that developed it. Basically, ACC is part of the ADAS applications and represents a functionality that can be used typically in complex scenarios such as highway environments and suburban roads with high driving speeds, overtaking vehicles and unexpected speed changes. In old cruise control systems, the driver could just set the wanted cruise speed and the vehicle will maintain it until the system will be enabled: This behavior persists independent of the environment situations and other vehicles motions. The only way for drivers to avoid accidents is to brake by itself when the front vehicles decrease their speeds.

This simple application has been improved during the years as much as possible, including an adaptive system that allows to maintain the safety distance changing autonomously and dynamically vehicle speed and direction according to the road conditions. A typical approach for ACC functionality is based on a simple scheme that sets the frontal vehicle closest to the car as the target to follow. Optical sensor (Sotelo et al. 2004), laser, and radar (Panaturak et al. 2008; Kim et al. 2009) are usually employed to perform this task. In order to percept and measure vehicles distances and the speeds toward the subject direction, the sensors are typically placed in the frontal part of the vehicle.

The perception system has to be able to replace the driver, operating in an autonomous way: It can operate without communicating with other vehicles or the driver, but it has to remain an assistance system; thus, it must let the control to the driver whenever he wants. Furthermore, it has to manage critical situations, for instance, in low free spaces and high speed environments, offering safety maneuvers. The current ACC research and development is focused on the prediction of these traffic conditions to provide a well-timed reaction in the presence of dangers.

The first step for ACC system development is the vehicle classification according to their relative positions. Three principal categories can be considered:

- Overtaken vehicles
- Nearby vehicles
- Medium-far distance vehicles

Basically, different regions index different methods to identify vehicles. In the *nearby* region, it is possible to see only a part of the vehicle from a frontal sensor (camera, laser or radar) because usually the car keeps all the field of view. In this case, vision techniques such as symmetry and shadow detection are deprecated.

In the *overtaken* region, vehicles reach high speed and only a part of them is visible; thus, an image motions analysis is more efficient than the other vision methods mentioned before.

Finally, the detection in the *medium and far* region is more simple because the vehicles are completely captured from the frontal cameras and lasers sensors. These regions are the most important for the ACC application because they delimit the area in front of the subject, where the objects' monitoring and identification are crucial operations for the correct functioning.

Moreover, lateral regions need to be monitored in order to promptly recognize warning situations, enhancing vehicle perception and improving the system capabilities.

Further consideration must be taken for the system reaction time and its reaction space: Every system has a specific working rate (i.e., a real case of study could be around 10 Hz) and a vehicle detection range (The minimum distance depends on the preceding vehicle position), currently it could be around 5–60 m up to 80 m with a tracking system. Considering this detection range and a comfort deceleration (3 m/s^2), it is possible to compute, for the automatic cruise control, the maximum speed difference between the vehicle and its preceding one as:

$$D = \delta V(t_l + t_a) + \frac{1}{2} \frac{\delta V^2}{a_c} \quad (43.1)$$

where D is the distance to reach preceding vehicle, δV is the speed difference, t_l is the localization time, t_a is the actuation time (so the relation $t_a - t_l$ is a reaction time), and a_c is the comfort deceleration.


Considering t_l equal to 0.3 s and t_a equal to 0.2 s, the maximum speed difference can be calculated as 60 km/h.

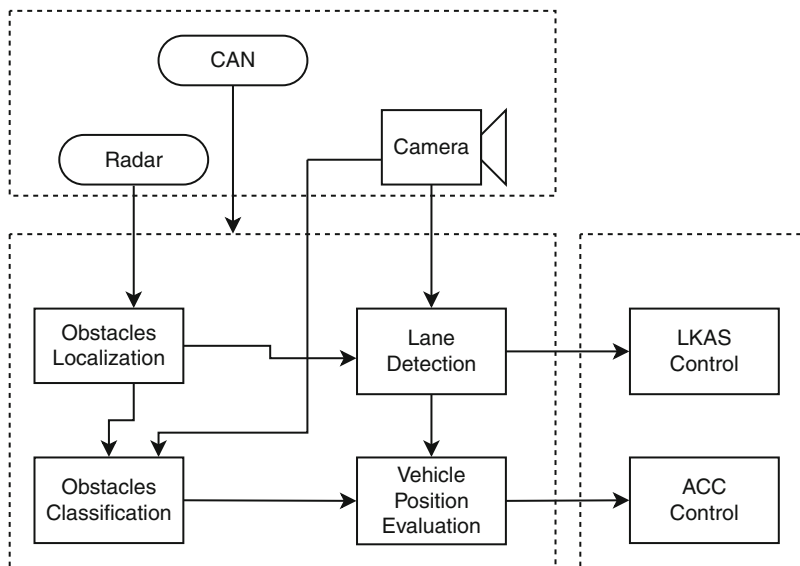
In case of frontal collision warning, the maximum speed difference to avoid the crash can be computed using the formula 2:

$$D = \delta V(t_l + t_a) + \delta V t_r - \frac{1}{2} a_c t_r^2 + \frac{1}{2} \frac{(\delta V - a_c t_r)^2}{a_s} \quad (43.2)$$

where t_r is the reaction time and a_s is the maximum deceleration for collision avoidance that a driver can issue. Considering t_r equal to 1 s and a_s equal to 8 m/s^2 , the maximum speed difference can be computed to be 80 km/h.

2 Systems

Described in  Fig. 43.1, there is a complete ACC system scheme based on passive and active sensor fusion with camera and radar sensors. In this case, passive sensor should be used to localize obstacles and their classification will be provided by both sensors data. Information about lane position with respect to the vehicle is used to control the *Lane Keeping Assistance System* (LKAS) and the information about preceding vehicles is used to control ACC. Active sensors are used to improve the system detection in front of the vehicle, because in that region, a total accuracy with a strong fail tolerance is needed.



■ Fig. 43.1

A scheme of a complete adaptive cruise control system with lane keeping system

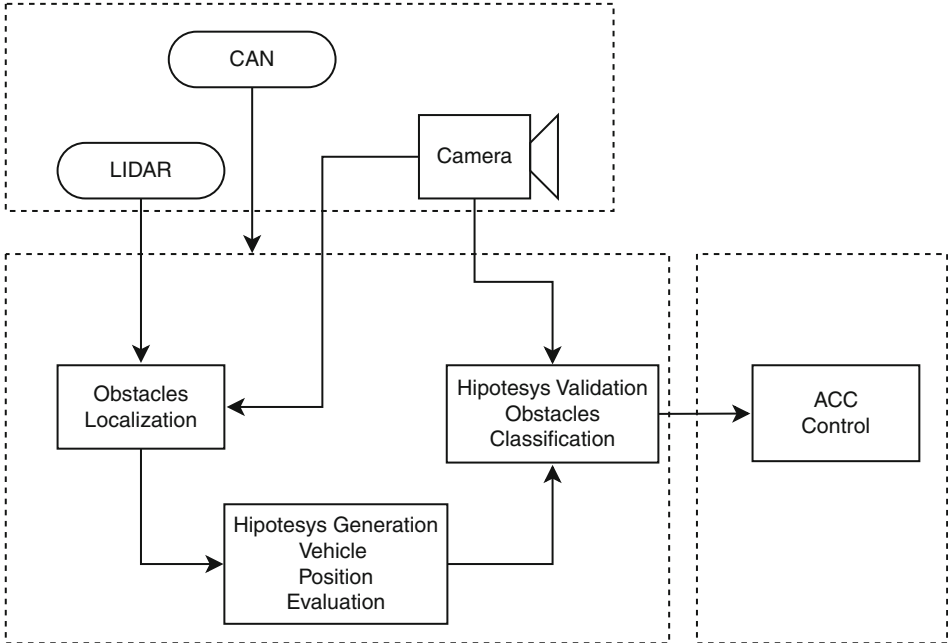
Otherwise it is possible to use a different set of sensors to implement an ACC system. Currently, LIDAR (Light Detection And Ranging) sensors are preferred in ADAS applications to obtain better result exploiting their high resolution and accuracy. Thus, it is possible to generate hypothesis of vehicles using these active sensors (Lüke et al. 2007; Hofmann et al. 2000) in order to restrict the passive sensor searching area, as showed in

► Fig. 43.2.

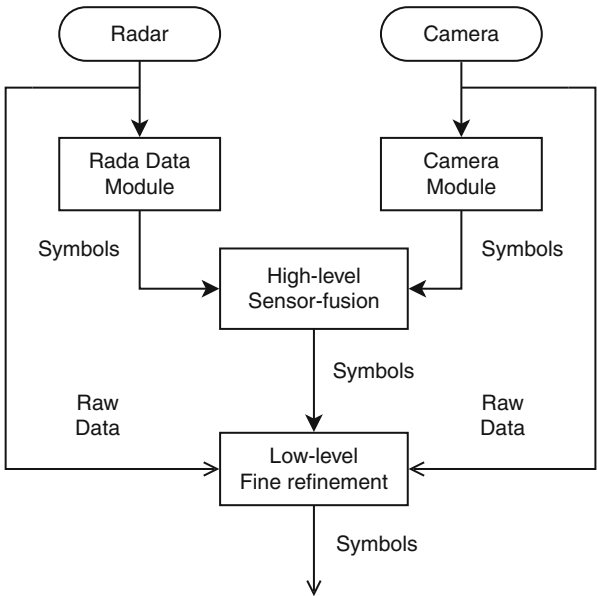
Most recent approaches are using active sensors to detect obstacles (such as radar and laser-scanners) in order to provide directly the distance measurement without any additional computation that can add some error to final results. The use of optical sensors for ACC system allows to spread the normal field of view of the system (up to 360°) improving the perception performance: This configuration can be used, in fact, to track overtaken cars and moving vehicles; moreover, it can support all ACC-related applications such as lane detection, sign's recognition and, additionally, the pre-crash warning system.

3 Algorithm Overview

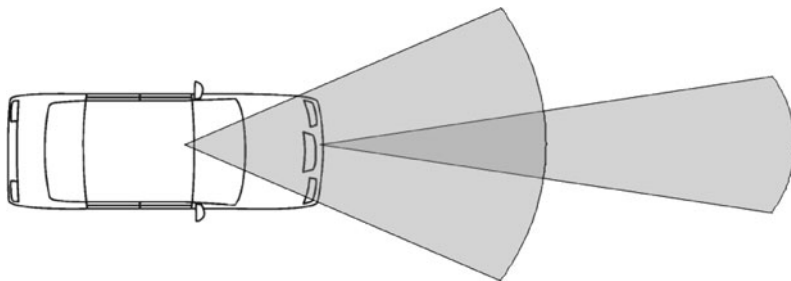
A possible algorithm, based on the system presented in ► Fig. 43.1, is shown in ► Fig. 43.3: It uses lane information to improve ACC and limit the vehicle's region of search. Basically, in order to obtain a more robust result, it is possible to refine high level data fusion with some low level data fusion. In this case, it is important to consider that low level fusion uses raw data, that are pure and unmodified data, while high level fusion uses some results



■ Fig. 43.2
A scheme of a complete adaptive cruise control system with LIDAR sensor



■ Fig. 43.3
Main flow chart



■ Fig. 43.4

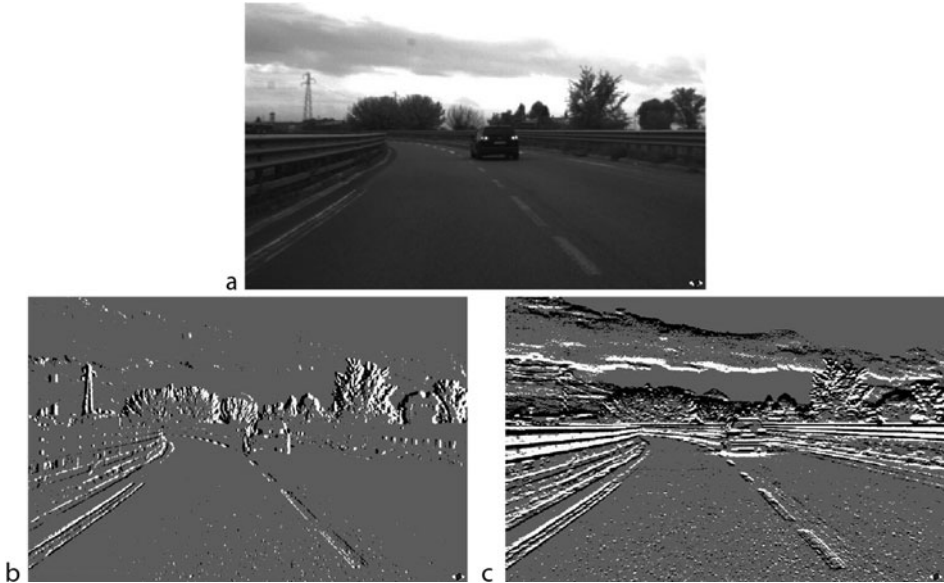
Sensors field of view

of processing phase like tracking data or the aggregation of raw data. High level data fusion can be used for ACC applications in simple environments, where the sensors are not affected by noise. To improve the performance in complex scenarios, low level fusion is a better choice for image processing: Raw data can be used to remove false detections obtaining better results.

Different algorithms can be used for on-board systems (as that shown in ● Fig. 43.1) exploiting the high accuracy of the active sensors with respect to passive ones: In this case, obstacle detection can be based on laser perceptions, focusing in a narrow band ahead area in order to achieve good results with low computational costs. Thus, vision perceptions can validate all obstacle candidates without exploring the whole image. Moreover, the vision information can be used to extend the perception FOV to the lateral sides and out of the Laser FOV (● Fig. 43.4). This algorithm allows to increase the computational efficiency during the images acquisition, saving time for any driver reaction needed. The proposed approach is based on two fundamental steps: The first one, called HG (Hypothesis Generation), performs hypothesis generation for the localization of all possible vehicles, while the second step, called HV (Hypothesis Validation), is an ad-hoc vision algorithm to verify the previous hypothesis.

3.1 Vehicle Detection

This section will present all methods used by the ACC systems showed in ● Figs. 43.1 and ● 43.2 to percept and identify vehicles. Initially a vision algorithm will be explained, then a clustering algorithm for laser-based detection will be discussed. Various approaches can be used for vision-based vehicle detection, the technique described here is based on the use of a single camera and it employs knowledge-based methods: Some discriminative vehicle features (such as symmetry, shadow, geometrical features) are extracted and processed in order to verify the presence of a vehicle.



■ Fig. 43.5

Original image (a). Vertical edges image (b). Horizontal edges image (c)

Edges. The first step for vehicle detection algorithm is the edge detection: Considering the frontal vehicles shape approximated as a “U,” specific patterns formed by two vertical edges and a horizontal one are searched in the images. Either vertical or horizontal edges are detected using the Sobel filter: Since the vehicle almost maintains the same position in different frames, only stable edges are considered. This allows to remove all noise contributions introduced by the background such as guard rail, trees, signs painted on the road, etc. Then all valid edges are binarized using a fixed threshold.

► [Figure 43.5a](#) shows an example of input image and its corresponding vertical (► [Fig. 43.5b](#)) and horizontal (► [Fig. 43.5c](#)) edges. For rear vehicle views, many horizontal and vertical structures are visible, such as shadow under cars, trunks, rear glasses, and lights, and a little bit less for vehicle roofs that have sharp edges due to a similarity between illuminated roof and background.

Symmetry. It is also possible to use symmetrical information to recognize vehicles. Images of rear and frontal vehicles are usually symmetric with respect to the vertical axis. In order to avoid some noise added to the symmetrical information, it is better to filter out any homogeneous areas.

Lights. During nighttime, the approach based on pattern analysis is not always reliable: The edges could not be well defined because of the low illumination and the shadow under the vehicles is not present. To cope with these problems, morphological operations are applied to binarize the source image in order to filter blobs that represent lights; then a light detection algorithm is used to evaluate the match between detected coupled lights

and well-known vehicle model. To perform this task, the algorithm has to consider a set of light features:

- Shape of single light: Usually two lights on the same vehicle have to be of the same shape. Information about light aspect ratio tolerance and area could be used.
- Size: Vehicle distance is known (i.e., laser or radar data are available or a calibrated camera is used), it is possible to filter out lights that are too big or too small.
- Distance between the two lights: It is limited to a range.

LIDAR-based method. The raw data can be transmitted from LIDAR sensors to a control unit (ECU) through the vehicle CAN bus, where these data are managed in a Cartesian coordinate system as 3D word coordinates. Then a clustering algorithm aims to transform a set of scanning points into a set of segments by a hierarchical agglomerated clustering algorithm. Different techniques can be used, depending on the data type coming from different LIDAR sensors. A general bottom-up approach can consider each single point as a cluster to merge successively into larger clusters. For the clustering operation, it is important to select a correct *distance measurement*, which will determine how much two points are closer to each other; this value also influences the cluster shape because some points could be closest to one point in some direction than other points from other direction. This *distance measurement* has to be evaluated considering the LIDAR sensor's attributes such as accuracy and spatial resolution; moreover, a distance function has to be considered according to the type of obstacles to detect.

A vehicle model could be represented by a rectangular shape with two long lines for modeling the left and right sides, and two short segments to model the rear and the front side. An important aspect to consider for this model is that generic vehicles can be sensed from a single plane laser-scanner from different positions and different orientations of vehicle model. Depending on the spatial resolution of the sensor and obstacle orientation, vehicles can be showed by the LIDAR as two segments (at least one segment). Anyway, whether sensor has low spatial resolution, due to a low level information, a single segment has to be model for any vehicle orientation; so, only a vehicle presence hypothesis could be done with good result. On the other hand, with high precision, more processing could be done to understand the vehicle orientation and enhance the next phases of vehicle detection.

4 Conclusions

This chapter outlined a practical example of vision-based system using computer vision technique and data fusion procedures. In the first part, parameters used by ACC such as region of interests, reaction times, and vehicle detection range have been introduced. A section briefly describes two real cases of study, and features of these systems are discussed. Furthermore, descriptions of algorithms applied to such systems are showed, using video data as a hypothesis generator or validator. The last session shows pros and cons in using video data in sensor fusion for ACC applications.

References

- Hofmann U, Rieder A, Dickmanns E (2000) Ems-vision: application to hybrid adaptive cruise control. In: Proceedings of the IEEE intelligent vehicles symposium, 2000, vol 4, Dearborn, pp 468–473
- Kim D, Moon S, Park J, Kim H, Yi K (2009) Design of an adaptive cruise control/collision avoidance with lane change support for vehicle autonomous driving. In: ICCAS-SICE, 2009, Fukuoka, pp 2938–2943, Aug 2009
- Lüke S, Komar M, Strauss M (2007) Reduced stopping distance by radar-vision fusion. In: Valldorf J, Gessner W (eds) Advanced microsystems for automotive applications 2007. VDI-Buch, Springer, Berlin, Heidelberg, pp 21–35
- Pananurak W, Thanok S, Parnichkun M (2008) Adaptive cruise control for an intelligent vehicle. In: IEEE international conference on robotics and biomimetics, ROBIO 2008, Bangkok, pp 1794–1799, Feb 2009
- Sotelo M, Fernandez D, Naranjo J, Gonzalez C, Garcia R, de Pedro T, Reviejo J (2004) Vision-based adaptive cruise control for intelligent road vehicles. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, IROS 2004, vol 1, Sendai, pp 64–69, Sept–Oct 2004

44 Vision-Based Blind Spot Monitoring

Elena Cardarelli

Dip. Ing. Informazione, Università di Parma, Parma, Italy

1	<i>Image Acquisition</i>	1072
2	<i>Vehicle Detection</i>	1074
2.1	Pattern Analysis	1074
2.2	Optical Flow Estimation	1076
2.3	Features Extraction	1077
2.4	Features Tracking	1079
3	<i>Algorithm Overview</i>	1083
4	<i>Conclusions</i>	1087

Abstract: These sections introduce a vision-based system designed for monitoring the area that a driver cannot see from exterior mirrors, usually referred to as *blind spot*.

This is a challenging task that requires to discriminate from vehicles and background when both are not static and also to cope with usual automotive problems like camera vibrations and oscillations.

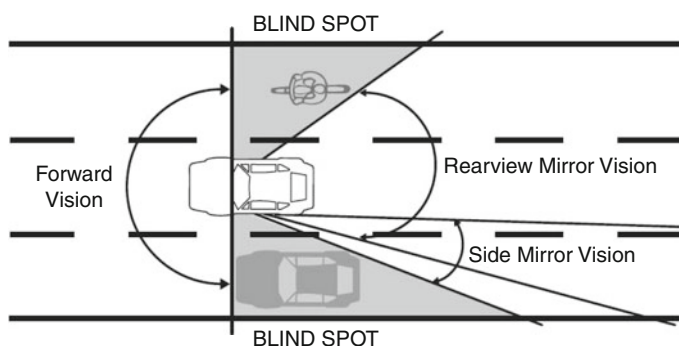
The development of ADAS has involved an improvement of the safety on the road, ensuring vehicle reliability and supporting driver for accidents preventing, as seen in the previous chapters. Particularly, some studies have been focused on the assistance during lane change, monitoring the area covered by the blind spot, which drivers are unable to see in the exterior mirrors (● Fig. 44.1). If the system detects an overtaking vehicle in the dangerous zone, visual and acoustic signals warn the driver about the risk of collision.

Radar technology and laser scanner are sometimes employed for the lane changing assistance, although they do not guarantee always a complete representation of the scene: When the vehicle is pitching down or up, the laser beam hits the ground or points to the sky, making data acquisition worthless. Moreover, in complex scenario, like urban road traffic, the radar has limited functionalities because its reliability depends on the radar cross section of the object to identify. For this reason, vision data fusion is necessary to provide high-resolution performance.

A common choice for the blind spot monitoring is the using of a camera mounted on the wing mirror (● Fig. 44.2) in order to extend the driver's view.

1 Image Acquisition

The use of a camera mounted under the side-view mirror increases the system complexity: The operability scenario is not static and different elements such as camera angle, perspective deformation, and camera vibration have been considered because they may affect the system performance. Moreover, when the camera roll angle is not null, the acquired images are rotated (● Fig. 44.3).



■ Fig. 44.1
Blind spot areas



■ Fig. 44.2

Example of camera placed under the side-view mirror of the VisLab car



■ Fig. 44.3

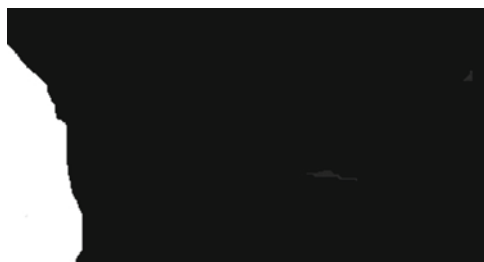
Rotated images acquired with a not null camera roll angle

To cope with this problem, images have to be transformed: With a rotation, this effect may be removed, but the relation between camera parameters and images becomes incongruous. For this reason, image rectification is performed: In this case, camera parameters are acquired and modified in order to obtain a null roll angle; then through the homographic transformation, the new parameters are assigned to the camera. The results of images rectification are shown in ► [Fig. 44.4](#).



■ Fig. 44.4

Examples of image rectifications to remove the camera roll angle



■ Fig. 44.5

Example of mask used to remove the vehicle where the camera is mounted on and wrong values

In order to remove the region occupied by the vehicle where the camera is mounted on, a gray level mask is applied to the image; an example is shown in ● Fig. 44.5: Only the images pixels corresponding to the black zone are considered useful information.

2 Vehicle Detection

The first step for the vision-based lane changing assistance is the detection of vehicles in the critical area; to perform this, task pattern analysis (VOLVO Technologies 2007) or optical flow estimation (Batavia et al. 1997) could be employed.

2.1 Pattern Analysis

In the blind spot application, pattern analysis is based on the detection of vehicle discriminative parts: Signature, lights, wheels, bumper, and plate could be searched in the acquired images in order to verify, with the support of a classifier, the presence of

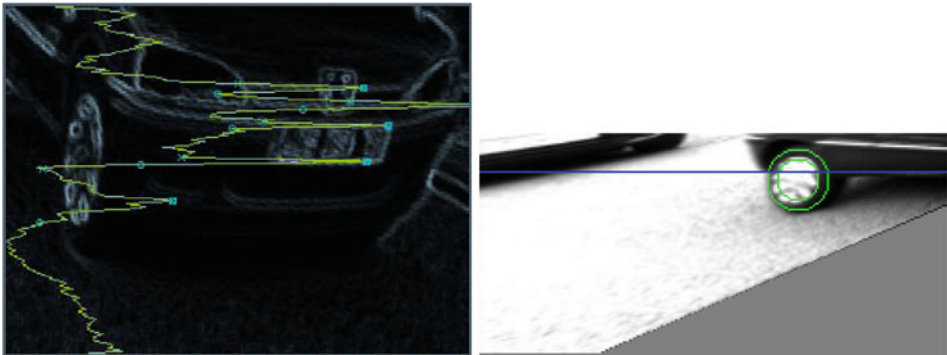
a vehicle. Also gradient information could be useful for this task. Moreover, tracking algorithms are employed to improve the reliability of the results.

When the overtaking vehicle is closed to the camera and its frontal part is occluded by the presence of the vehicle where the camera is mounted on, template matching could be used (► Fig. 44.6).

The drawback of these approaches concerns the generalization of the detection for all type of vehicles providing, at the same time, low computational costs: since cars, motorcycles, and trucks have different features, the recognition step could be complex, affecting real-time performance.

Moreover, in the pattern analysis, the shadows under the vehicles could be considered in order to delimit the region of interest where vehicle are detected using distinctive information, like symmetries, edges, or shape (► Fig. 44.7).

In this case, the processing involves low computational costs, but the system performance is limited because the extraction from a single image of information about object position is not accurate.



■ Fig. 44.6

Wheel and frontal vehicle detection based on template matching



■ Fig. 44.7

Using the shadows for vehicle detection

2.2 Optical Flow Estimation

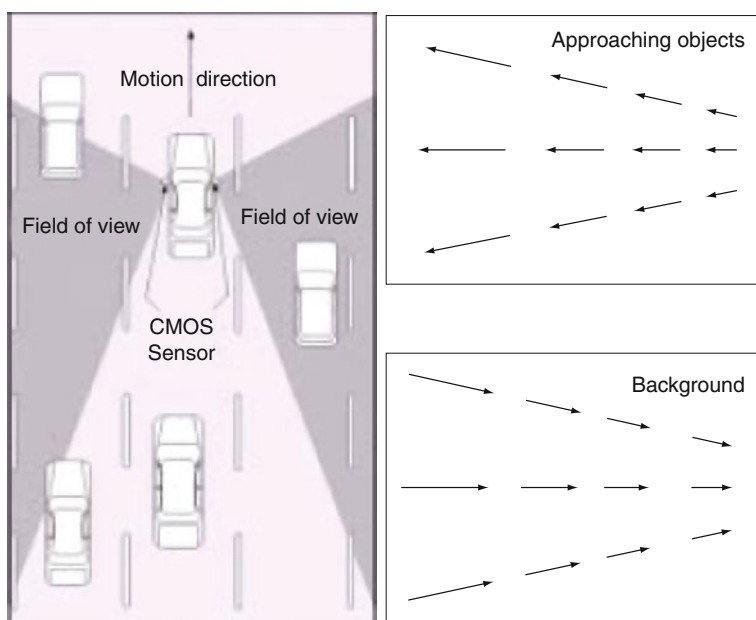
Many approaches (Batavia et al. 1997; MVT Ltd 2004) are based on the evaluation of the object moving with respect to the camera's one: Since the camera is mounted on the side-mirror, the vehicles' detection have to be performed in a dynamic scenario (🔗 Fig. 44.8).

To recognize an overtaking vehicle all directions of image elements are processed: If an object is moving closer to the camera, it is considered as an overtaking vehicle; otherwise, it is a background element.

The optical flow estimation is a common choice to represent the apparent speeds distribution generated by the objects motions, and usually, the estimation is supported by a tracking algorithm that allows to search in the current frame the position of previously detected objects in order to estimate their speed and direction. Moreover, to increase the accuracy, the optical flow estimation could be combined with edges detection (Sun et al. Mar. 2006; Mae et al. 1996), color information or Kalman snakes.

Since the optical flow of the background elements is similar to the moving objects' one, the studying of the FOE (Focus Of Expansion) could be a possible solution to increase the reliability of the results and limit, at the same time, the computational costs. In this case, the road plane is estimated in the image in order to delimit the region of interest for the vehicle searching.

Either dense or spread optical flow analysis may be performed: In the first approach, all image pixels are considered as contribution for the motion estimation; in the



■ Fig. 44.8

Description of the scenario and the motion of the objects in the image



■ Fig. 44.9

Background modeling: dynamic areas subtraction

second case, only some distinctive features are processed. In ► Fig. 44.9, an example for overtaking vehicle detection based on sparse optical flow analysis and eigenvalues processing is shown: In the image, dynamic and static areas are separated, then sparse optical flow is used to subtract dynamic region from the background in order to make the detection robust to camera shocks and vibrations.

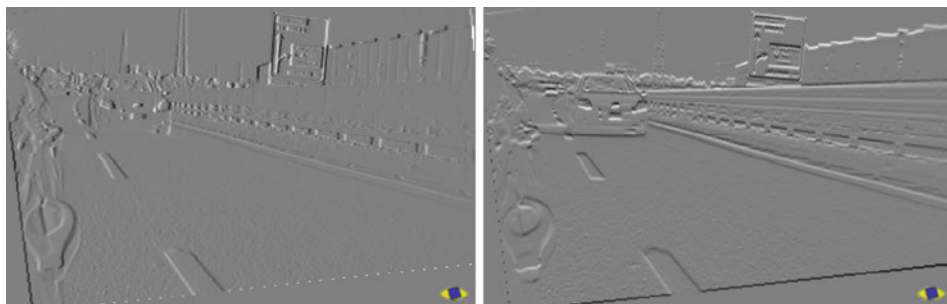
An important aspect of the optical flow is its versatility: it allows to discriminate among static objects, elements with the same direction of the vehicle where the camera is mounted on and vehicles that follow an opposite direction with respect to the camera's one.

For the optical flow estimation, it is firstly important to extract from the images a set of discriminative features: These points are tracked among difference frames, in order to determine their motion with respect to camera's one. Then the features are classified according to their direction, in order to select and pack together all points belonging to a possible overtaking vehicle.

2.3 Features Extraction

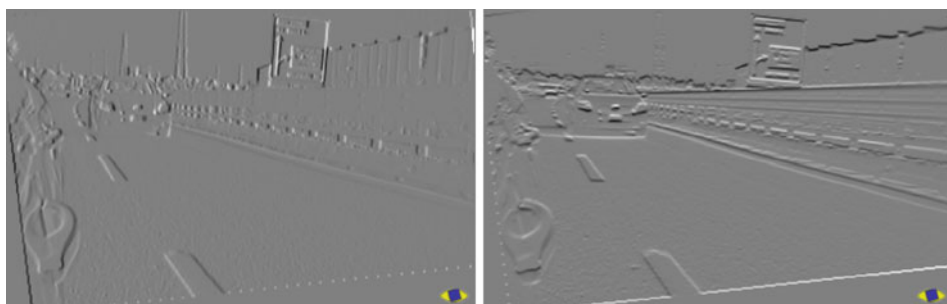
Every system based on features tracking strictly depends on the feature's extraction algorithm: If a feature detected in a frame does not appear in the next one, the input data for the tracking are unreliable. An important aspect to consider in the feature extraction is that in the image, some points could not be tracked: In shape and color homogeneous areas, where there is not texture, it is impossible to determine exactly where a previously detected point appears in the current frame; thus for some group of pixels, it is impossible to track their motion. Moreover, an object edge allows to determine the element motion only along the direction perpendicular to the edge. Instead in the presence of a corner, there is a significant brightness change that allows motion tracking along both axes.

Feature extraction can be performed directly processing gray level images or using edge-based algorithms; in this case, all points with maximum curvatures or edges crossing points are considered. In ► Figs. 44.10 and 44.11 are shown the images resulting by the application of two common methods for edges enhancement: gradient operator and Sobel filter.



■ Fig. 44.10

Vertical and horizontal edges enhancement obtained by gradient operator



■ Fig. 44.11

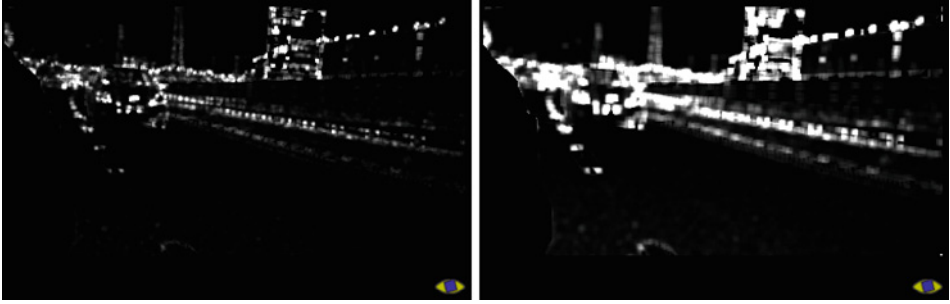
Vertical and horizontal edges enhancement obtained by Sobel operator



■ Fig. 44.12

Shi-Tomasi feature extractor

In order to guarantee the reliability of the results, in automotive application, the principal aspects to be considered for the choice of the best features extractor are: temporal stability, accuracy on feature detection, and computational costs. Some examples of feature extractor are shown in ► [Figs. 44.12](#) and ► [44.13](#).



■ Fig. 44.13

Harris corner detector using respectively a 3×3 and a 5×5 window. Starting from the edges images, all significant corners are detected. By increasing window dimension, the corner detection becomes more accurate, but the computational cost is higher

2.4 Features Tracking

Feature tracking is based on the generation of a correlation among objects in different frames: it focuses on the detection in the current frame of previously detected features, in order to determine their motion, and estimate their future position.

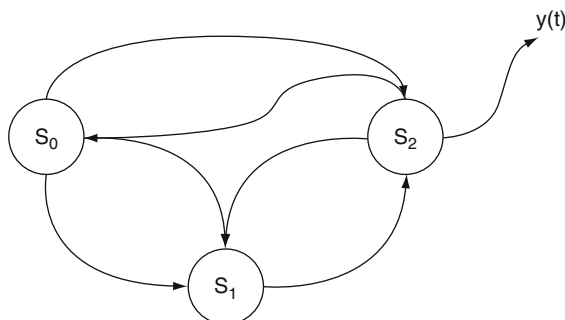
The information to track consists of color distribution or edges, even if the use of color images is not a common choice because illumination and weather condition could affect the system performance. The choice of the suitable feature tracker depends on different aspects: the tracking model to use (statistics, volume, etc.), the dimension of the features space (2D or 3D), the type of acquisition system (monocular, stereo, trinocular), and the camera mobility (static or not).

Tracking could be focused on the extraction of specific features, useful to identify vehicles' parts such as wheels: In this case, a model of the object to track is defined using a priori knowledge. In practice, model-based tracking exploits distinctive information about the elements to track or their motions to limit the set of region of interest to process.

Otherwise tracking may be performed after the detection, using as input all the region of interest, without any information about high level knowledge: In this case, the system is based on motion detection and tracking of the detected areas in order to determine the trajectories of the moving objects. This choice allows to provide generalization about the application scenario.

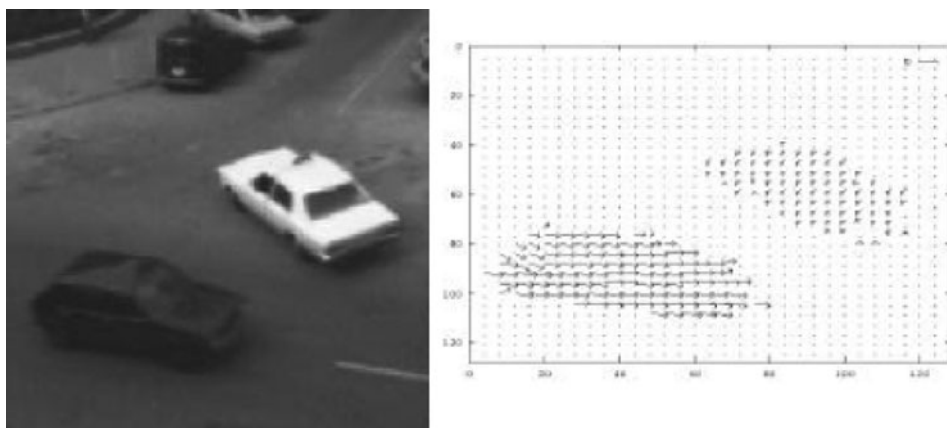
An example of non-model-based tracking approaches is the Hidden Markov Model, a probabilistic technique based on the analysis of time-discrete series: A basing structure is shown in ● Fig. 44.14. A generic state is connected with a certain probability to other states; the interconnection parameters may be estimated using feed forward techniques such as the Baum–Welch algorithm. The features to track could be points, lines, or two-dimensional blobs.

Another method for feature tracking could be based on optical flow estimation (● Fig. 44.15): In this case, the information about objects motion is important to detect



■ Fig. 44.14

An example of three states HMM, $y(t)$ is the observation obtained by each state



■ Fig. 44.15

Optical flow estimation for features tracking

scene dynamics and correlate spatial information with time variation. Starting from image points, the optical flow estimation provides a two-dimensional approximation of the 3D world points motions. To perform the tracking, all features with a uniform motion are packed together. The processing is repeated in different frames in order to determine in the images all elements with a similar motion and speed. Finally, elements' trajectories are determined exploiting borders and center information of the detected region of interest. A priori knowledge about objects to track may improve the performance.

Differences between optical estimation and real motion may affect the performance: Since only the apparent motion may be extracted from the processed images, to obtain a quantitative evaluation of the scene, some additional assumptions have to be done, for example, brightness changes, objects specific features, and the relation between world points and 2D projections motion. Template matching is another possible tracking

approach: It is based on the localization of a specific region tracked using correlation evaluation. Usually, the area of interest is selected in order to make its detection easier: It involves the processing of specific image points called corner points (i.e., where the brightness gradient is high) and their neighborhood. The advantage of this technique concerns its flexibility. The template is frequently updated in order to make the tracking robust to scene variation: If the template in the current frame is significantly different from the previously detected one, a new template is defined according to the new region of interest; in the presence of slight template changes, a mean template version is processed to cope with small noise deviation. In the template matching, the correlation techniques could be region based or feature based according to the information to track.

In [Fig. 44.16](#), an example of features' tracking based on template matching is shown: The gray level intensity in a neighborhood around each detected feature is tracked between different frames; moreover, the position with respect to the vanishing point is considered in order to verify that the features' motion is compatible with an overtaking vehicle motion. In particular, in a typical dynamite scene, the direction of the overtaking vehicle's features is opposite to the vanishing point motion.

To evaluate the features' motion with respect to the vanishing point, it is possible to calculate the angle between the line that crosses the previous detected feature and the vanishing point, and the line defined by the current feature and the previous detected one. This angle is equal to:

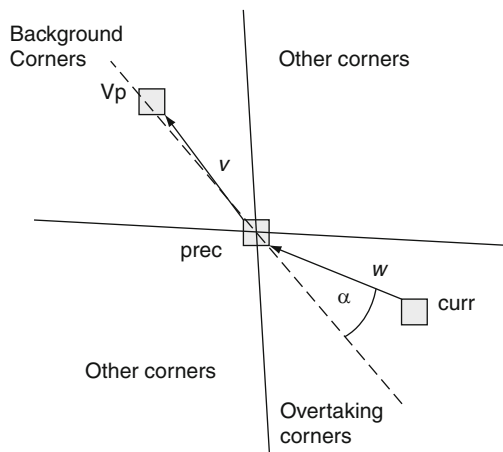
$$\cos\alpha = \frac{v \cdot w}{\|v\| \|w\|} \quad (44.1)$$

[Figure 44.17](#) shows a possible approach for the determination of the background and overtaking vehicles features according to their direction: V_p is the vanishing point, $prec$ is the previous detected point, and $curr$ is the current feature. α is the angle



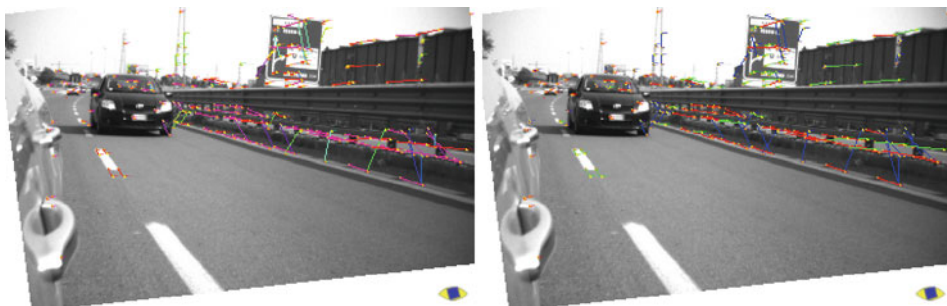
■ Fig. 44.16

Template matching for features tracking



■ Fig. 44.17

Analysis of the angles between matched features




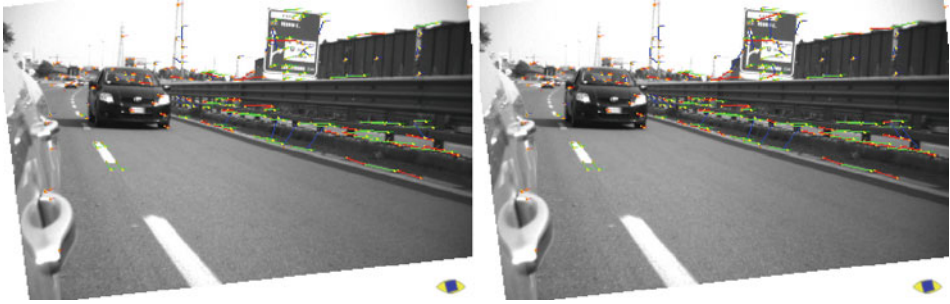
■ Fig. 44.18

Features classification according to their included angle

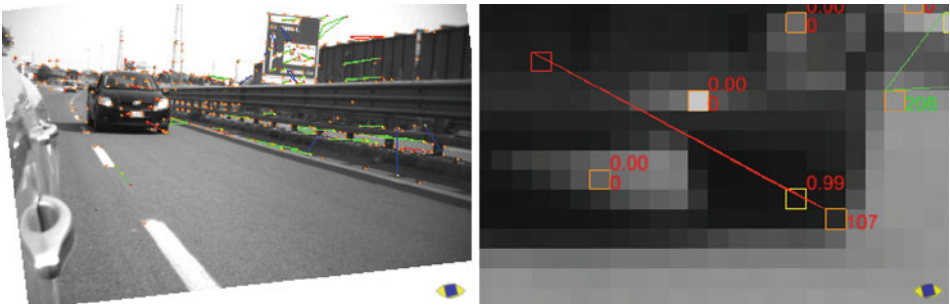
between the lines defined by the vectors v and w . The matched corners are classified according to their motion as:

- *Overtaking corners*: They represent the overtaking vehicles, with a positive speed greater than the speed of the car where the camera is mounted on. The features pairs associated to this category have the $\cos\alpha$ positive and greater than a certain threshold.
- *Background corners*: They are associated to the background elements and to the objects with a negative speed; in this case, the features matches have $\cos\alpha$ negative and lower than a specific threshold.
- *Other corners*: They represent all other image elements with a motion not compatible with the previously described categories.

In  Fig. 44.18, the tracked features are plotted with different colors, according to the included angle; in particular, the right image shows the classification results: Overtaking



■ Fig. 44.19
Reliable features tracking and noise filtering



■ Fig. 44.20
Tracking of the feature's history

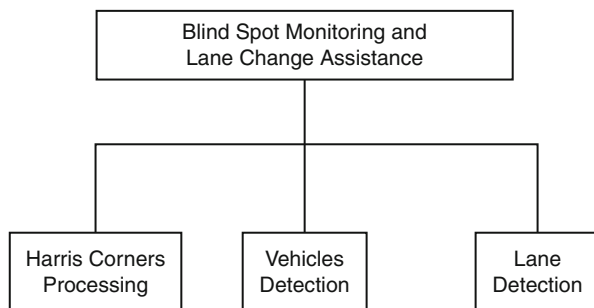
vehicles are represented by red segments, background elements are enhanced with green lines, and all other objects are plotted with blue segments.

To track the pairs of features, best matches between the current points and the previously detected ones could be considered, associating a specific vote to each couple of features. Moreover, if the vote is not greater than a default threshold, the relative features pair is discarded in order to remove noisy contribution. The results obtained with this noise filtering are shown in [Fig. 44.19](#).

To improve the results, it is possible to maintain the history information for each tracked features pair: For example, the position of the first instance, the current, and the previous features coordinates may be considered to count the number of frames on which that feature's pair has been correctly detected. In this way, it is possible to calculate the “age” of each instance and use this value to further filter the results. In [Fig. 44.20](#) are shown the results of the described features history tracking.

3 Algorithm Overview

A general scheme for blind spot monitoring is shown in [Fig. 44.21](#): The top module defines a high level algorithm that interacts with three independent blocks, that



■ Fig. 44.21

Example of blind spot monitoring scheme

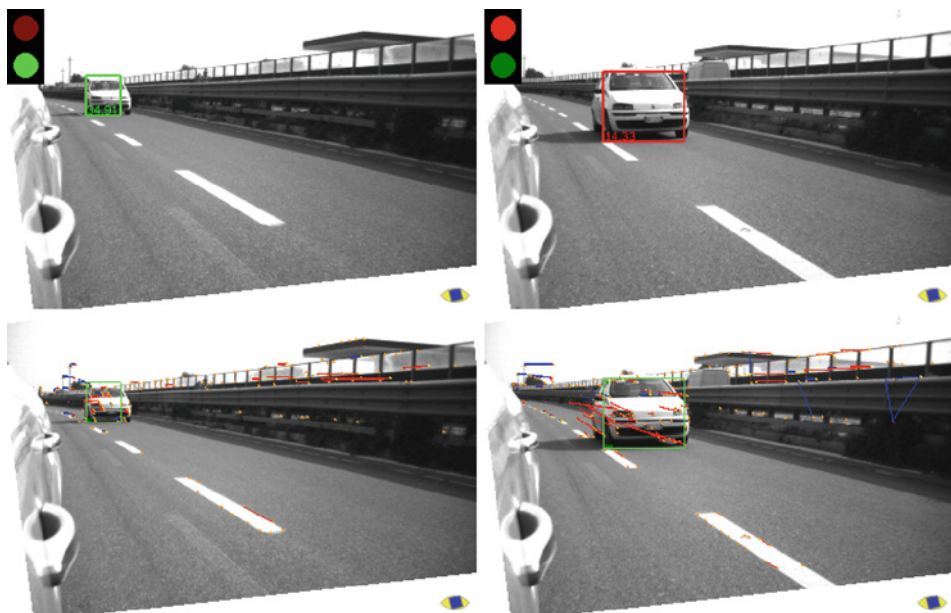
implement low level routines. These blocks operate simultaneously and communicate their outputs to the top level module. All the results are fuse together by the top level module in order to determine the appropriate behavior.

For each image acquired by the camera, Harris corners are extracted and tracked in order to detect all moving objects in the scene. According to pattern matching approaches, overtaking vehicles are recognized. Simultaneously lane detection is performed, allowing the high level application to warn the driver when an overtaking vehicle will occupy the lane closer to his car; moreover, with multi lanes detection, it is possible to evaluate their intersection in order to improve the FOE estimation and obtain a value that does not strictly depend on the calibration parameters.

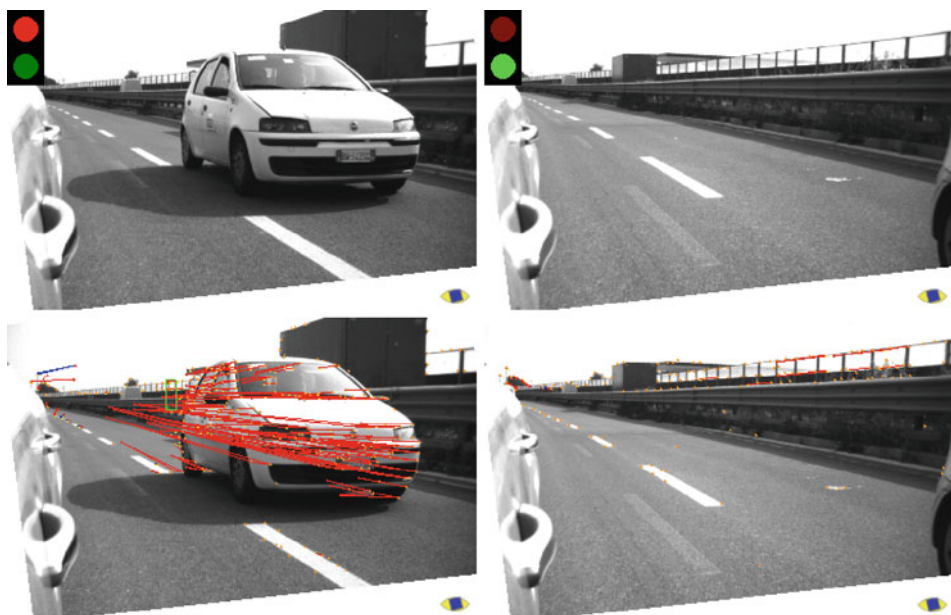
To identify all dangerous overtaking vehicles, the high level module selects among all possible candidates only the cars within the limits defined by the second lane, if it is detected. To reduce the number of false positives, the selected elements have to be tracked in different frames and they have to be described by a significant number of Harris corners. If the distance of the overtaking vehicle from the camera is lower than a certain threshold, an acoustic and visual signal is activated to warn the driver about the dangerous situation.

An example of functioning is shown in ● Fig. 44.22: The left images do not represent dangerous situations because the vehicles detected in the second lane are far from the camera, thus the warning signal is not emitted. Vice-versa the scenarios in the right images described a critical situation: The vehicles detected are close to the camera; therefore, the red light is on and the car is delimited by a red bounding box to enhance its dangerousness.

When the acoustic/visual signal is activated, the blind spot area is monitored to warn the driver about the presence of vehicles in that region. Therefore, the signal remains enabled if the number of Harris corners relative to overtaking vehicles is greater than a specific threshold; otherwise, it is automatically switched off. In ● Fig. 44.23, the presence of a vehicle in the spot area involves the enabling of the warning signal (red light on) in the right image; when the vehicle leaves the dangerous zone (left image), the signal is deactivated.



■ Fig. 44.22
Example of lane change assistance system

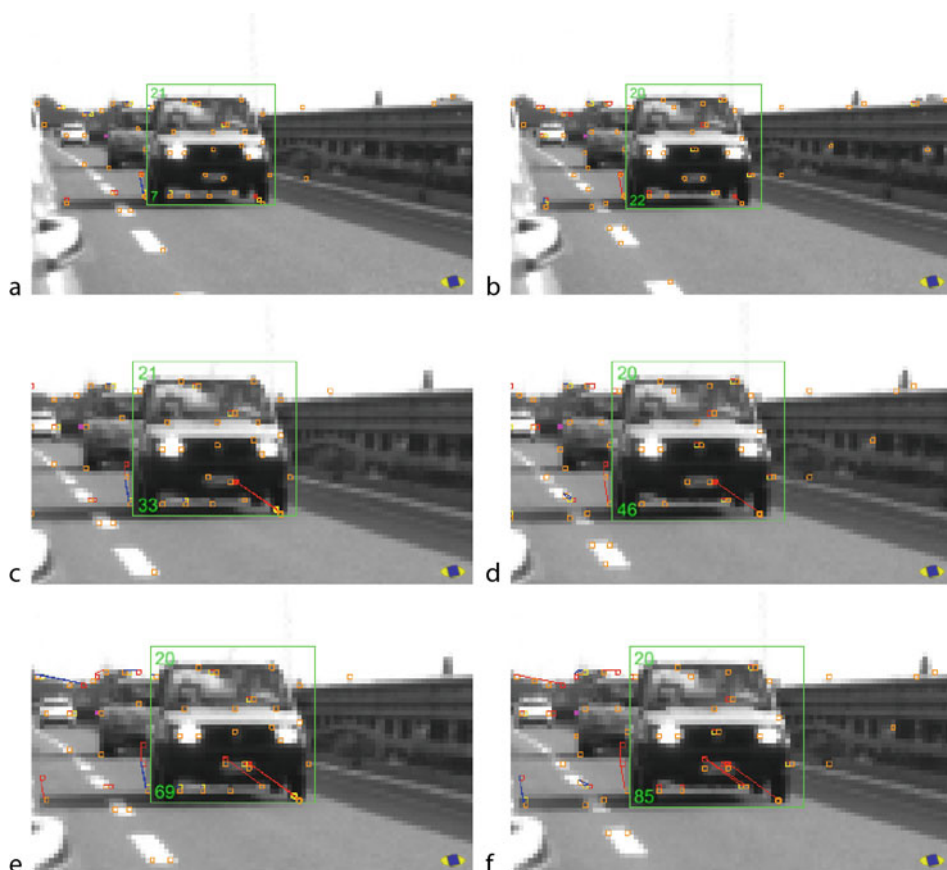


■ Fig. 44.23
Example of vehicle detection in the blind spot area

In the system, different thresholds are defined in order to provide the reliability of the results:

- *Dangerous Distance*: It determine if the spatial limits within a lane change maneuver is considered dangerous.
- *Min Corner Vehicle*: It sets the minimal number of features needed to consider a candidate as an overtaking vehicle.
- *Min Age Vehicle*: It is the minimal tracking history for valid vehicles.
- *Min Dangerous Corner Blind Spot*: It defines the minimal number of corners in the blind spot area to consider the region occupied by a vehicle.

In **Fig. 44.24** are shown the Harris corners tracked with the previously described approach, varying the dimension of the window use to detect the maximum values in the Harris image. By increasing this dimension, the detected corners are more stable, thus the tracking algorithm provides better performance.



■ Fig. 44.24

Harris corner tracking with different windows dimensions: (a)–(b) 5 pixel, (c)–(d) 7 pixel (e)–(f) 9 pixel

4 Conclusions

The implementation of a blind spot monitoring system, based on computer vision algorithms, for lane change assistance is an important issue for providing safety on the road. If the lane change maneuver is performed while an overtaking vehicle occupies the dangerous zone (i.e., delimited by a critical distance), an acoustic/visual signal is activated to warn the driver. A particular attention has to be reserved for the camera calibration phase to guarantee the reliability in the measurements of the overtaking vehicles distance. Moreover, with the image pre-processing, it is possible to provide image rectification and improve the vehicles detection performance. A blind spot monitoring system may be based on a layered architecture where different modules perform three principal tasks: vehicle identification, dynamic objects extraction, and lane detection. Several methods may be used for the detection of the overtaking vehicles: pattern analysis, optical flow estimation, features extraction, and tracking. A possible approach for overtaking vehicles detection could be based on their shadows analysis; the Harris corners may be used to discriminate between static and dynamic objects.

References

- Batavia PH, Pomerleau DA, Thorpe CE (1997) Overtaking vehicle detection using implicit optical flow. In: Proceedings of the IEEE international conference on intelligent transportation systems'97, Boston, USA, pp 729–734, Nov 1997
- Mae Y, Shirai Y, Miura J, Kuno Y (1996) Object tracking in cluttered background based on optical flows and edges. In: Proceedings of the 1996 international conference on pattern recognition (ICPR'96), vol 1–7270. IEEE Computer Society, Washington, DC, USA, pp 196–200
- MVT Ltd. (2004) Mobileye N.V. Blind spot detection and lane change assist (BSD/L-CA). <http://www.mobileye-vision.com>
- Sun Z, Bebis G, Miller R (Mar. 2006) On-road vehicle detection: a review. IEEE Trans Patt Anal Mach Intell 28(5):694–711
- VOLVO Technologies (2007) Blind spot information system (BLIS) by volvo. <http://volvo.com>

Vehicular Communications Systems

Scott Andrews

45 Vehicular Communications Requirements and Challenges

Scott Andrews

Systems Engineering and Intellectual Property Management in the Automotive, Mobile Computing, and Communications Domains, Cogenia Partners, LLC, Petaluma, CA, USA

1	<i>Vehicle Communications Overview</i>	1093
2	<i>System Concepts</i>	1093
3	<i>Geographic Context</i>	1094
4	<i>Temporal Context</i>	1096
5	<i>Reliability Context</i>	1100
5.1	RF Signal Level	1100
5.2	Multipath Effects	1101
5.3	Hidden Terminal Effects	1103
5.4	Density and Bandwidth/Congestion Effects	1106
6	<i>DSRC Overview</i>	1109
6.1	Communicating Entities	1110
6.2	Channels	1110
7	<i>Mobile Terminals</i>	1111
7.1	Embedded Vehicle Terminal	1112
7.2	Aftermarket Vehicle Terminal	1113
7.3	Portable Consumer Electronic Terminal	1114
7.4	Infrastructure Terminals	1114

8 *Example Implementations* 1115

8.1 Signage 1115

8.2 High Priority Signage 1116

8.3 Lower Priority Signage 1117

8.4 Probe Data Collection 1117

8.5 Traffic Signal Violation Warning 1118

9 *Conclusion* 1120

Abstract: Outlines the primary requirements and constraints applicable to vehicular communications. Describes key operational issues and summarizes typical applications. Provides overview of Dedicated Short range communications (DSRC) and describes typical implementation configurations.

1 Vehicle Communications Overview

The past decade has seen tremendous progress in the evolution of wireless data systems. While a large portion of this progress has been in the development of wide area networks to support consumer demand for mobile Internet access and messaging, significant technical progress has been made in communications systems to support safety and mobility applications. While some of these applications can be addressed by wide area consumer wireless systems, the unique nature of the vehicular applications imposes many demands that render conventional wireless solutions inadequate or overly complex.

2 System Concepts

Conceptually a vehicle communications system is relatively simple. Mobile terminals, typically embedded in vehicles, are configured to exchange messages with other mobile terminals, and these same terminals are able to exchange messages with fixed infrastructure systems. Generally the objective of vehicle-to-vehicle exchanges is to communicate vehicle state information that may then be used by various safety applications to either inform the driver of hazardous situations, or to control the vehicle itself. Vehicle-to-infrastructure exchanges typically involve a wider range of motivations, including access to remote services, receipt of roadway related safety information (e.g., hazards and intersection information), and various local transactions such as toll payments.

In general, for the purposes of this chapter, the communicated information has some relation to the location of the mobile terminal. While conventional wireless internet access is well known and widely available in consumer electronic devices, the internet services available are typically not directly related to the immediate location of the terminal. Although clearly there are exceptions to this, this chapter will not focus on this type of internet communication, and instead will concentrate on peer-to-peer communications related to safety in the immediate vicinity of the vehicle, and the communication of information from fixed sources related to the immediate vicinity of the vehicle. The reason for this focus is that from a practical perspective, the most important safety threats to a vehicle are in the immediate vicinity of the vehicle. Mobility threats are generally wider in nature, but still centered on the current location of the vehicle, and possibly its intended route.

The environment in which this communication is carried out is challenging since the vehicles are typically moving at high speed relative to the infrastructure, and may encounter each other only very briefly. This communication is further complicated by the fact that there are millions of vehicles on the road, and at any given instant a vehicle

may be in the immediate vicinity of several hundred other vehicles, yet, while one might be able to physically see these vehicles, each of them is, at least at the start of the communications process, anonymous (that is, it has no network address or other identity). The only common thread between the vehicles is that they are at that particular location on the road. Establishing a unique addressed communications exchange with each vehicle is problematic since each communicating terminal (fixed or mobile) must establish several hundred simultaneous sessions, and, since the vehicles are moving, these sessions will be constantly changing. Simply learning the network identities of each terminal in this environment is likely to consume the entire time of the encounter. The larger the radio range the less the session turnover, but conversely, the greater the number of sessions that must be managed (since larger range implies a larger number of communicating vehicles in range). So, at root, the vehicle communications challenge is one of efficiently communicating with other mobile terminals and other fixed services in a constantly evolving system configuration within a relatively constrained local area.

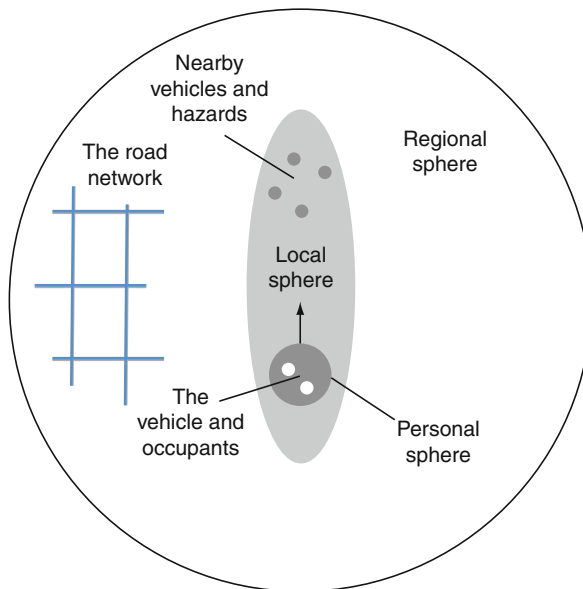
3 Geographic Context

Mobile users have a variety of communication needs that cover a relatively large breadth in terms of distance. An inherent feature of the mobile communications task is that some operations are local in nature, while others are may be regional. And, in some cases, communications may be limited to sending messages between devices inside the vehicle. These different communication spheres arise because of the wide variety of information that concerns the vehicle and the vehicle user. For example, a mobile user is unlikely to be concerned about an unknown mobile user miles away, but may be highly interested in a mobile user a few meters away. In contrast, this same mobile user may also be concerned about traffic several kilometers away, and traffic is an aggregated result of many mobile users, but not of any particular one. These “spheres of concern” can be roughly arranged as shown in ● Fig. 45.1. This figure describes several spheres of concern which relate to different physical or geographic regions around the mobile terminal. For example, a mobile terminal in a vehicle may be influenced by a *personal sphere* wherein information is passed through an interface to and from the vehicle in which the terminal is currently placed. This same terminal will also seek to exchange information with other terminals and with other services in the immediate vicinity of the terminal. From a practical perspective, this *local sphere* can be assumed to extend to about the distance from the vehicle that it would require to stop. For example, when a vehicle is stopped the potential threats to that vehicle are generally in very close proximity (perhaps a meter or so away). However, at speed the *local sphere* will extend away from the vehicle since any given threat will require some distance of travel to avoid. In a still larger context, the *regional sphere* may relate to the overall road network, including all of the other vehicles not in the local sphere.

Generally, different communications media are optimized to support each of these spheres. So, for example, the personal sphere might be served by a wired interface, or possibly by a personal area network, such as Bluetooth. In this application the vehicle, the

occupants, and the mobile terminal are in close proximity to each other for long periods of time, so a short range wireless system, or a wired connection, are appropriate. Conversely, in the regional sphere, the elements of concern are large-scale things, such as the road network, or the weather, that cover a wide area. Regional elements may affect, for example, decisions about which route to take, or where to find a particular type of food. This type of information may be useful to many vehicles across the region and, at the same time, may be of no interest at all to most vehicles across the region. In this way it is very much like the World Wide Web where the users of any particular piece of information are dispersed. As a result, the regional sphere is best served by a wide area communication system such as the many cellular data systems available for wireless Internet access.

The local sphere has several unique qualities. First, it is usually in motion relative to the regional sphere, and the elements of concern (e.g., roadway hazards, and other vehicles) may be entering and leaving the local sphere continuously. Second, its extent depends, to first order, on the vehicle speed. The faster the vehicle is traveling, the more the local sphere must extend in the direction of travel. Third, the elements of interest within the local sphere are generally small scale. They may be, for example, a pothole in a particular lane, or animals migrating across the roadway. And, fourth, the local sphere of a particular vehicle will intersect the local spheres of many other vehicles more or less randomly. From a practical perspective, there are simply too many local elements to catalog for an entire region, so while it might be possible to use a regional communications system to provide information about local elements, it would not be particularly practical. First, the system would need to catalog potentially millions of local elements across the entire region, and second each local




■ Fig. 45.1
Communication spheres

terminal would need to continuously query the system as it moved in order to learn what local elements are in the (new) immediate vicinity. While this approach might work conceptually, it does not scale well since information of interest to each vehicle must be individually selected and sent to each vehicle, and as each vehicle moves the process must be repeated. At substantial vehicle penetration, and in areas of high hazard density, the number of requests and responses would be staggering. In addition, since the vehicles may be moving relatively quickly, the update rate and latency requirements for such a regional system would be effectively unachievable.

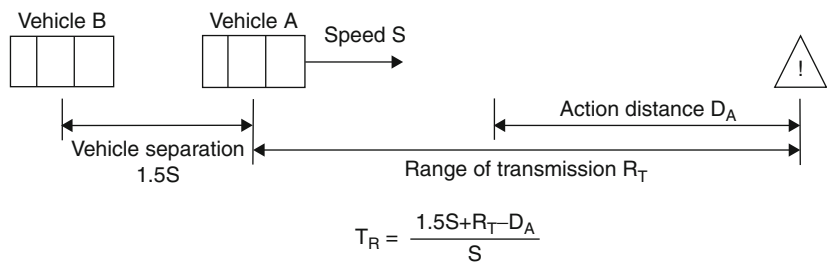
An alternate approach is to use localized broadcast transmissions in a non-networked communications environment. Not only does this approach avoid the need for each vehicle to request local information from some regional source, it also eliminates the need for a vehicle to establish any network presence at all. This approach is highly scalable, and has the added benefit of being structurally anonymous. In this model, instead of each terminal having a network address, messages are simply transmitted in the local area where they are relevant, and any other terminals in that area can receive them. The mechanisms for this type of communication are discussed more fully later in this chapter.


4 Temporal Context

Because of their different geometric characteristics, different wireless safety and mobility applications have different temporal requirements. For example, a message warning about an upcoming curve in the road must be transmitted frequently enough to assure that approaching vehicles receive the message early enough to take appropriate action (i.e., slow down). This is illustrated in  Fig. 45.2 below.

Here,

- T_R is the maximum time interval between message transmissions.
- R_T is the range of the transmitter in meters.
- D_A is the distance in meters at which action must be taken (this is the minimum distance at which the message must be received).
- S is the vehicle speed in meters per second (typically the 85th percentile speed for the road).
- The vehicle separation time is assumed to be 1.5 s.



 Fig. 45.2
Message timing diagram

Using this scenario, in the best case, a vehicle receives the message at R_T , and assuming the action distance D_A is less than R_T , the vehicle can take appropriate action. If this occurs, and a second vehicle is following at 1.5 s, then the worst case is that the message does not repeat until this second vehicle is D_A away from the hazard. The message must thus repeat at this time or earlier to assure that the second vehicle will receive it with sufficient time to take action. Obviously, if R_T is less than D_A the system will not work, so this also provides an indication of the minimum range requirements for various applications.

► **Table 45.1** below provides typical intervals between message transmissions for various applications. Here we have estimated the typical 85th percentile speeds for each application situation, a 150-m transmission range, and certain desired actions on the part of the application. For example, for crossing path vehicles, and highway and city street hazards, we have assumed that the appropriate action is to reduce speed by 25%, for traffic signals and stop signs we have assumed the vehicle is to stop. For approaching path vehicles we have assumed that the vehicle must reduce speed and change lanes (in some cases, increasing speed and changing lanes may be the safer action, but for illustrative purposes we have used this action).

This table provides some interesting insights about application design. For example, in the traffic signal situation, we have assumed that the message contains the current signal phase, and the time until the next phase change. This means that with a single message, the vehicle application can identify and characterize the traffic signal hazard, and, in the worst case scenario, this message will be received at the distance the vehicle requires to stop. So, despite the fact that the light may change to red after the message is received, a properly designed application will have already determined the point at which this event will occur (based on the information in the previously received message). An alternative approach would be to simply send a message with the current signal phase (i.e., no additional information about the length of that phase). Under this scheme the message would need to repeat at the average following time interval (the time between vehicles), which is

■ **Table 45.1**
Maximum application message repeat interval

Application	Typical speed (kph)	Action distance (m)	Maximum message repeat interval (s)
City Street Hazard	40	5.5	14.5
Highway Hazard	100	34.5	5.6
Stop Sign	40	12.6	13.9
Traffic Signal Phase and Timing	60	28.5	5.8
Traffic Signal Phase Only	60	28.5	1.5
Crossing Path Vehicle	60	12.4	9.8
Approaching Path Vehicle	200 (combined)	211.6	0.4

typically 1.5 s. Here, a lead vehicle might receive the message that the signal phase had changed from green to yellow just late enough that the application would assume that the vehicle should continue through the yellow light (presumably the yellow phase would be set to allow this). The next vehicle behind would need to receive the message no later than 1.5 s after the prior message, since by that time it would be at the stopping distance threshold (i.e., at the action distance from the signal).

Of course, if the system has additional features, such as signal phase intervention for emergency vehicles, then the application timing will need to take this into account.

It is also important to note that since wireless communications is inherently unreliable because of variations in RF propagation, interference, and noise, it may be desirable to repeat the messages somewhat more frequently in order to increase the probability of reception.

Traffic density also includes two key factors that influence other roadside transmitter parameters. Specifically, the rate of traffic flow past the transceiver (which is determined by the expected number of vehicles per hour, and the average speed of those vehicles). These factors interact with the range of the transceiver to determine how many vehicles the transceiver must serve and how quickly it must serve them. This is illustrated in [Fig. 45.3](#). This figure shows the transceiver encounter time as a function of range and vehicle speed.

The box overlaid on the graph illustrates the encounters that are between 5 and 20 s. In general, encounters of less than 5 s are unlikely to be completed if there is any complexity to the transaction and/or if the service is remote (so that the transactions must pass over the backhaul to a remote server and then return). And, at the other end of the spectrum, most ITS applications are not expected to require more than about 10–15 s total, so there is no need to provide for longer encounters.

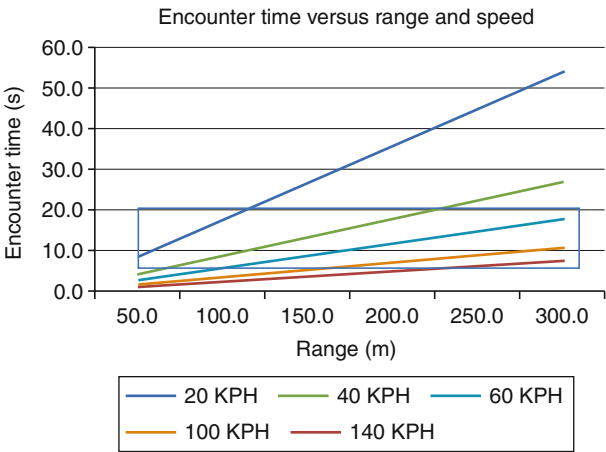



Fig. 45.3
Transceiver encounter time versus range and vehicle speed

There are two ways to reduce or control the effective range of the transceiver. One is to simply reduce the power so that a typical vehicle transceiver will exhibit a minimum packet error rate at a proscribed distance. This approach is not particularly accurate since different vehicles may possess different radio sensitivity and antenna gain. Lower power also often results in problematic radio behavior at the fringes of coverage. The better way to control the range is to limit the geographic scope identified in the security certificate. This is not a suitable approach for mobile transceivers, but if the transceiver is fixed, this approach allows the RF power to be high so that when an vehicle transceiver is inside the defined geographic region the packet error rate is very low.

From  Fig. 45.3, it is clear that if the average traffic speed is slow, it may be highly desirable to limit the range using either of the above methods. In some applications where the traffic speed may change substantially, for example, during rush hour, it may be advantageous to dynamically change the range so that the number of vehicles being served per second remains approximately constant.

The fact that the user vehicle population is moving also substantially limits the ability of a localized communications system to support traditional networking. Conventional wireless networking systems (e.g., 802.11 g) include an association process whereby each new wireless terminal joins the network. In this process the base station indicates the channel it is using, assigns an IP address to the terminal, and identifies the other terminals in the local network. This approach was developed for conventional local area networks (LANs) so that, for example, two computers can share files, or a computer can identify a printer and send files to it to be printed. In the roadway context, however, this approach is not practical. From the perspective of a single vehicle, other vehicles and roadside terminals are coming and going all the time, so if the system were to establish a conventional network, the members of the network would be inconstant flux. The same is true for a roadside terminal, which might act as a conventional hot spot. In this approach, the vehicle terminals would be in constant flux, and the system would spend an inordinate amount of time updating the network configuration. The concept of a network also suffers from ambiguity in situations where, for example, a group of vehicles traveling down the road have formed a network and then encounter another group of vehicles that have also formed a network. In this case the networks must either merge, possibly very briefly, or the individual terminals must become part of both networks (or possibly several networks) at the same time. Either of these situations represents a substantial network management problem that can consume more resources than the messages the network is intended to convey.

For this reason, most vehicle communication systems do not rely on the conventional network concept, and instead rely on broadcasting messages to whatever terminals happen to be in radio range. There are exceptions to this, for example, in situations where the vehicle terminals are serving as nodes in a regional network (e.g., an ad hoc mesh system), but it is not clear if these systems provide any significant advantage over the more efficient and more easily managed wide area cellular networks.

5 Reliability Context

Wireless communications is inherently unreliable because of a number of factors, and this reliability has a direct impact on the effectiveness of these systems. This is especially critical for safety applications since, obviously, if a safety message is not received, then the system fails to provide the safety benefit. Factors that impact message reliability are RF signal level, multipath fading, hidden terminal collisions, channel access congestion, and external interference.

External RF interference is typically managed through frequency spectrum regulations and transceiver design, although it is generally wise to also examine the local RF environment when siting a roadside transceiver.

5.1 RF Signal Level

RF power has a clear but complex relationship to the effectiveness of communications. Successful communications is based on the link budget, which accumulates the various gains, losses and noise sources between the transmitter and the receiver. The resulting signal-to-noise ratio, in the context of the radio modulation and detection scheme, determines the probability of correct detection of the received signal, and thus the reliability of the communication. While it is not necessary to go in to the details of this analysis here, it is useful to understand the role of RF power in communications and how that power may or may not affect the quality of communications.

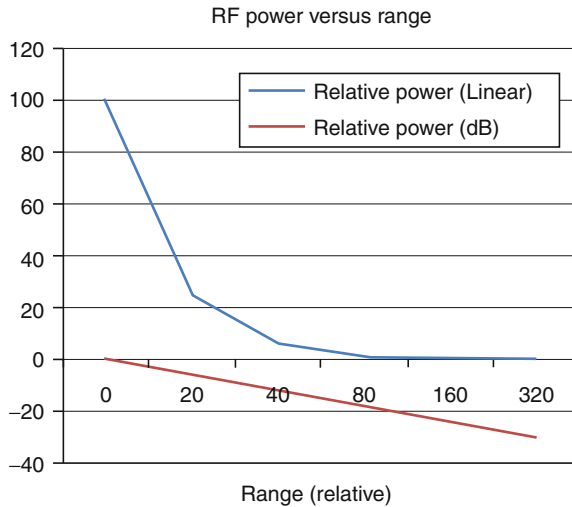
Radio power falls off as the square of the distance, so doubling the distance results in one fourth the power. On the other hand, if there is sufficient power to reliably communicate, there is (to first order) no harm in receiving more power from a transmitter (unless the power is so high it causes other issues, but this is generally not a major concern).

► [Figure 45.4](#) shows the change in RF power as a result of range.

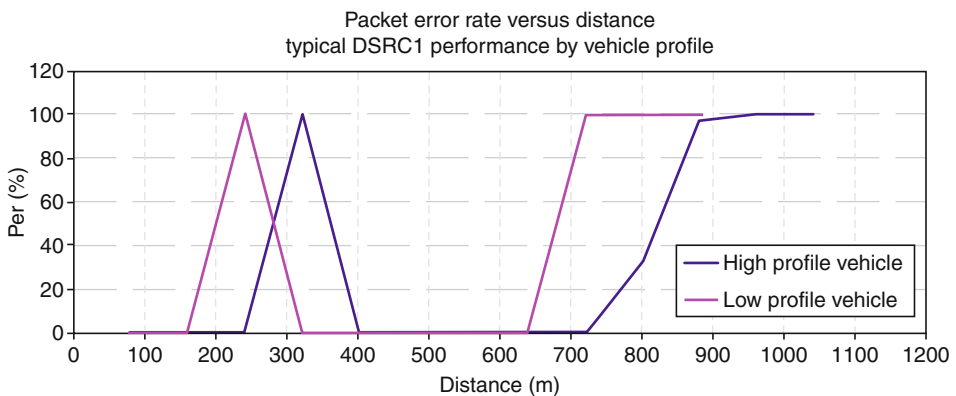
Careful examination of ► [Fig. 45.3](#) shows that for each doubling of the distance, the power falls by a factor of 4 (6 Decibels, or dB). A link budget analysis will identify the range at which the signal-to-noise level will not allow reliable detection of the radio signal. At this level the number of detection errors rises quickly, and the resulting packet error rate will also rise quickly. However, when the link is “closed,” that is, when there is sufficient RF power that the signal-to-noise level allows reliable detection of the radio signals, the error rate falls rapidly to near zero. Beyond this detection threshold, more RF power will not improve the communications. At this point communication errors are dominated by other influences.

This can be seen in ► [Fig. 45.5](#) below. This figure, taken from the VII Proof of Concept Testing Report (Volume 3a), shows the packet error rate for two different vehicle receivers as they move toward a roadside transmitter.

In this case the packet error rate falls very quickly from 100% (no communication) to effectively zero (good communication) within 50–100 m (in the 650–750 m range). Using the $1/R^2$ model described above, the expected change in RF power between 100% PER and 0% PER over this range is about 1 dB.



■ Fig. 45.4
RF power versus range



■ Fig. 45.5
Packet error rate versus range

The rises in packer error rate in the 200–300 m range are a result of multipath fading, which is described below.

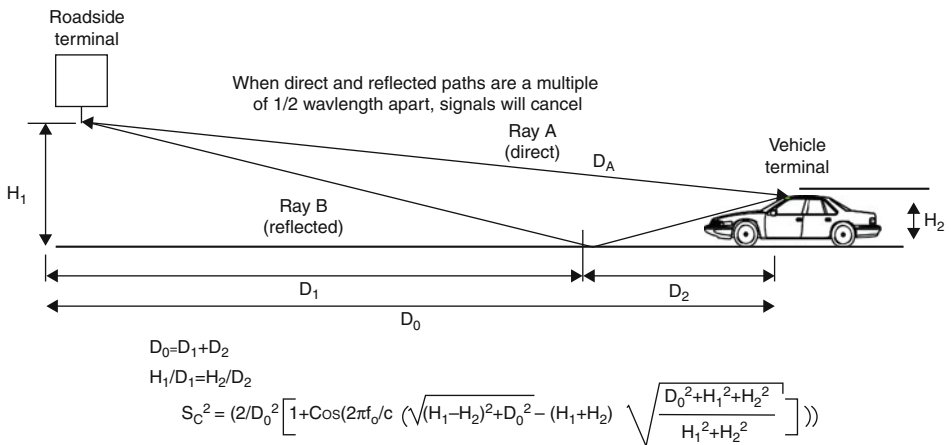
5.2 Multipath Effects

Typically radio systems are configured to provide a direct line-of-sight between communicating terminals. This is especially true as the operating frequency rises, which is typical of higher bandwidth digital communications systems. In general, higher frequency

RF signals do not bend along the curvature of the terrain and do not penetrate trees and other vegetation very well. However, these signals do reflect off of a variety of surfaces, especially relatively smooth surfaces such as roads and buildings. In general, reflected signals are unwanted because they interfere with the direct signal and can degrade the communications link. Under certain cases the reflected signal (with certain power levels and delays relative to the direct signal) can cause complete loss of signal packets and major degradation of communications. As shown above, when operating in a low signal-to-noise region, a small additional change in the signal level can effectively terminate communication.

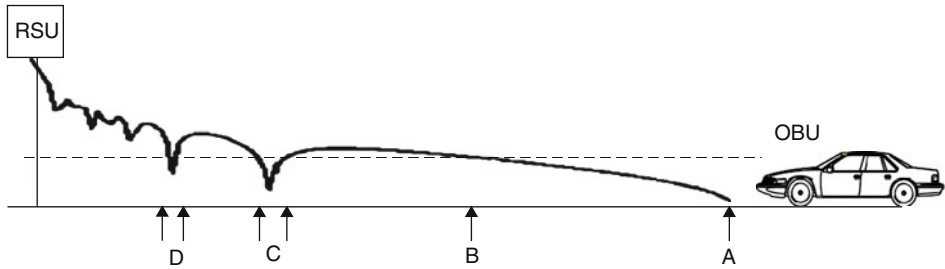
Multipath is typically described using the “two ray” model. This is illustrated in [Fig. 45.6](#) below. Here, a transmitter is located at one elevation and a receiver is located at another elevation, and both are separated by a distance D . The transmitter antenna is assumed to radiate RF energy in all directions. The primary RF path is the direct Ray A between the transmitter and the receiver. There is a second path in which Ray B reflects off the road (or some other specular surface) and then propagates to the receiver. The location of the reflection point is a function of the separation distance D , and the elevations of the two terminals. Since the path lengths between the two terminals are different for Ray A and Ray B, the phase of these two signals at the receiving terminal will be different. As the distance D changes, this phase difference will also change, and the resulting composite signal at the receiver will vary in amplitude. This variation is given by the equations in [Fig. 45.5](#).

[Figure 45.7](#) below illustrates the received signal strength as a function of distance. In the figure we have overlaid an arbitrary minimum signal-to-noise level (dashed line) at which the link budget is closed and reliable communications can be achieved. At point A the signal level is well below what is required for useful communications. Point B represents the location where the signal level is high enough that the system can



■ Fig. 45.6

Multipath two ray model



■ Fig. 45.7

Signal strength measured by OBU approaching transmitter

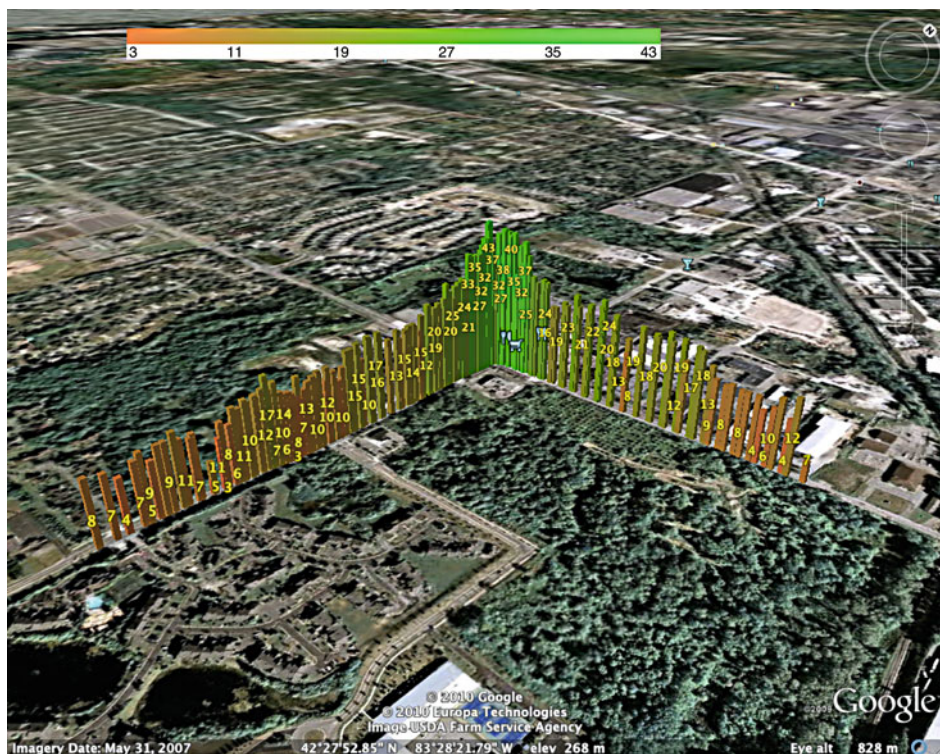
communicate. The signal rises (roughly following the $1/R^2$ relation described above) until the vehicle approaches point C. In this region the phase relationship between the direct and reflected rays (Ray A and Ray B) becomes destructive, and the signals begin to cancel each other. Once past this point the interference becomes constructive and communications can once again take place until the vehicle approached point D where the signals cancel again. As can be seen in the figure (or observed by plotting the equations), the composite signal level varies with a series of superimposed oscillations as the vehicle moves over very small distances. A few of these (in this example points C and D) are significant enough to terminate communication.

► [Figure 45.8](#) below shows the Received Signal Strength Indication (RSSI, an index related to signal strength in IEEE 802.11 communications) as measured by a vehicle receiver approaching a roadside transmitter. The figure clearly shows the periodic variations and nonuniformity of the received signal strength as the vehicle approaches the transmitter.

5.3 Hidden Terminal Effects

Vehicle based communications systems are inherently multiuser. In order to support multiple users these systems implement some form of medium access control (MAC). The MAC layer typically implements a mechanism to prevent users from transmitting on the channel at the same time, and thus interfering (Note: CDMA systems are designed to support multiple simultaneous users, and thus do not require that only one user transmit at any given time; however, they require non-overlapping codes, and this is problematic for a broadcast-oriented system, since each receiver must determine the codes being used, and with numerous codes corresponding to numerous vehicles, this would be excessively time consuming).

Most systems use some form of Carrier Sense Multiple Access (CSMA) wherein each transmitting terminal listens to the channel. If the channel is busy (i.e., another terminal is transmitting) it selects a randomly generated time interval. When the channel becomes clear (no other terminal is transmitting) it then waits for the selected time interval, and then begins transmitting. If another terminal starts transmitting before the time interval expires (because it either chose a lower random back-off interval or has been waiting



■ Fig. 45.8

Received signal strength versus range

longer), then the terminal in question stops counting down, and waits until the channel clears, whereupon it starts counting down again from where it left off. In this way, those terminals who have been waiting the longest are served first (because there is less time left in their countdown interval). In this way the transmitting radios in a given area are interleaved, and the probability of successful (non-interfering) communication is high.

CSMA is not perfect however. For example, it is possible for two terminals to select the same time interval and then transmit at the same time. This is situation occurs when there are a substantial number of terminals all seeking to transmit, or when two transmitting terminals are out of range of each other, the so-called “hidden terminal” situation.

If the transmission is unicast, that is the message is addressed to a specific terminal or network address, then the sending terminal typically expects some form of acknowledgment. These may be high-level acknowledgments as used in TCP/IP, or they may be very low level acknowledgments as used in UDP. If the acknowledgment is not received in this mode, the sending terminal will select a larger random duration interval, and try again.

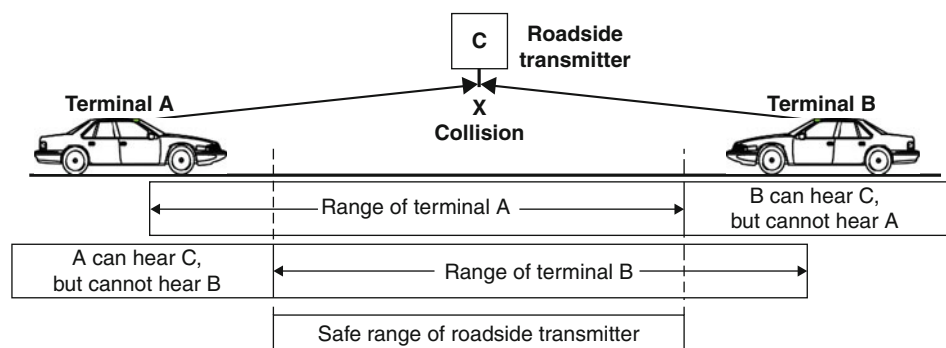
If the transmission is broadcast, then there is no specified destination terminal, and thus there is no acknowledgment. In this case the sender cannot know if the broadcast message was received by other terminals, or if it collided with another message

transmitted at the same instant. It is thus important to remember, when dealing with broadcast messages, that there is no way to know if others received it, so such applications should be designed with this in mind.

► [Figure 45.9](#) illustrates a typical hidden terminal situation. In this figure, vehicles A and B can hear the roadside terminal, but, because they are out of range, neither can hear the other. According to CSMA rules, each will listen to the channel, and, if the roadside terminal (or any other terminal in range) is not transmitting, each will hear no channel activity, so they will send their messages. The roadside terminal will hear both messages at the same time, and will be unable to separate them, so the transmissions will fail. Only when the vehicle terminals are within the region identified as the “Safe Range” will they be able to hear each other.

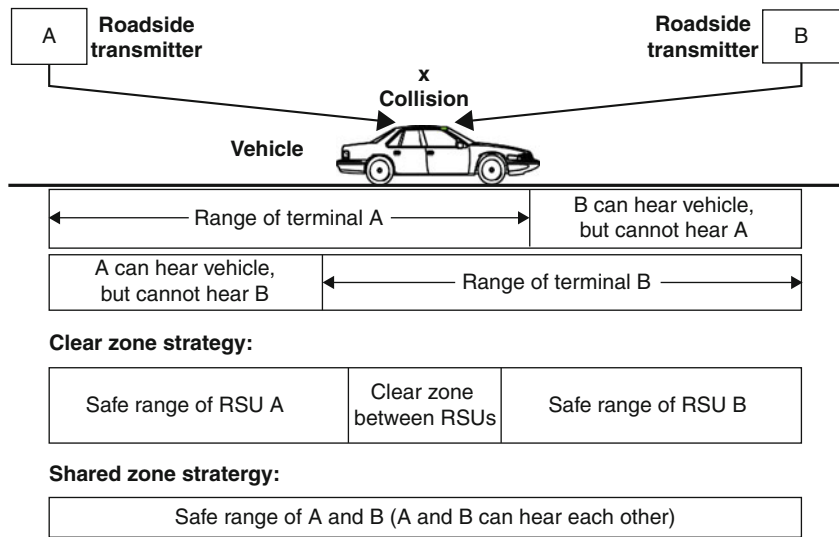
The complementary situation can occur with two Roadside Units (RSUs) and a single On-Board Unit (OBU), as shown in ► [Fig. 45.10](#). Here the RSUs cannot hear each other, and so they may send their messages at the same time. In the zone between the RSUs where an OBU can hear both RSUs, these overlapping messages will also fail. This situation may not be particularly problematic, since the OBU approaching any one of the RSUs will eventually be out of range of the other RSU, and thus messages from that RSU will be received without interference. To avoid this situation the RF power of the RSUs in this situation may either be limited so that there is a sufficient clear zone for the approaching OBU to interact with the RSU without interference, or the power levels should be high enough that the RSUs can hear each other and the CSMA scheme will prevent them from transmitting at the same time. These two schemes are denoted in ► [Fig. 45.10](#).

There are various other schemes to prevent this situation. One can, for example, use different channels to send message in the vicinity of a roadside terminal that when out on the open road, but, these schemes cannot avoid collisions in all situations, so it is important to design applications and transmission repetition rates with this issue in mind, in order to improve the reliability of the system.



■ Fig. 45.9

Hidden node situation (roadside terminal between vehicles)



■ Fig. 45.10
Hidden node situation (vehicle between roadside terminals)

5.4 Density and Bandwidth/Congestion Effects

Vehicle communications typically involves a large number of communicating entities. In this way it is somewhat different from other communication systems. As discussed earlier, setting up and managing a conventional network in the vehicle-roadway context is problematic because of the flux of network members, and the problem of overlapping networks. In addition, the potentially high level of vehicle density creates other challenges. The bandwidth of a given communications channel must be shared among all of the users within radio range, so the combination of the number of users and the bandwidth each requires sets an upper bound on the population that can be supported, and/or a lower bound on the communications bandwidth. This problem is aggravated by the fact that safety and mobility are continuous processes, so each terminal is likely to be using the channel regularly and on an ongoing basis. Conventional cellular systems are designed to support a large population of infrequent users, most of whom are using relatively low bandwidth voice communications, or very low bandwidth text messaging. Because of this, the emergence of multipurpose portable computing devices such as smart-phones has created substantial challenges for cellular carriers because the bandwidth usage models for these devices is significantly different from the conventional voice and text model. When all users are downloading music, or streaming video, the load on the network escalates, and the assumptions about sharing the channel bandwidth begin to falter. Vehicle communications is similar. While the vehicle messages are relatively short, the number of vehicles in a given area can be quite high, and the frequency of their messages can also be high.

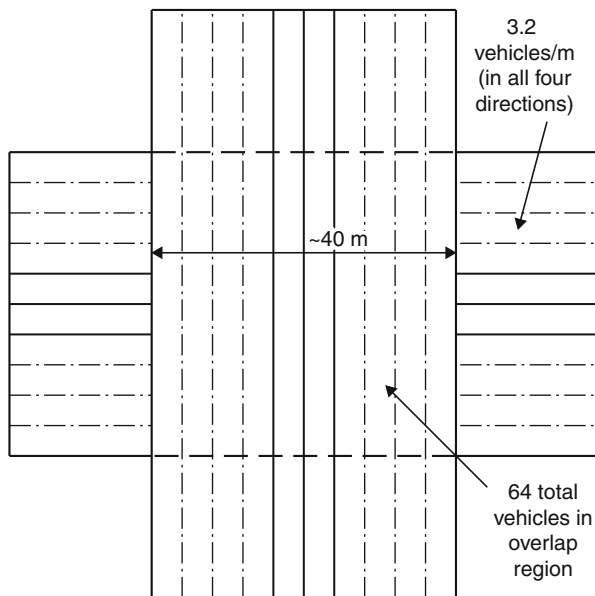
The number of terminals that must share the channel bandwidth is determined by the vehicle packing density and the size of the RF footprint for the channel. A typical vehicle is

about 5 m long, and a typical lane is about 4 m wide. On the road, vehicles are seldom closer than about one vehicle length, so the typical roadway vehicle density is about one vehicle per 10 m per lane, or 100 vehicles per kilometer per lane. A worst case packing situation might occur during rush hour, where an 8 lane freeway might be carrying 800 vehicles per kilometer. If two such freeways cross, the overlapping region (about 40 m²) will contain 64 vehicles, and each meter outside this range will contain an additional 3.2 vehicles. ► [Figure 45.11](#) illustrates this situation.

► [Table 45.2](#) shows the number of vehicles using the channel as a function of distance from the center of the overlapping zone together with the bandwidth per vehicle under several different assumptions about overall channel bandwidth, for a 500-byte message (a typical vehicle-to-vehicle message with security overhead).

As can be appreciated from the table, if all vehicles are using the channel simultaneously (as they might be if they are all sending vehicle-to-vehicle messages to each other), the overall bandwidth per vehicle is relatively low, especially as the range of the system increases.

An alternative way to understand the impact of shared channel bandwidth is to look at the situation where a channel access control system (e.g., CSMA) provides all of the channel bandwidth to each user when they have a message to send. In this situation, the optimal approach would have the system grant each user access to the channel in sequence until all users have sent their message. The performance measure then is the time before any one user can send another message. This is shown for various system bandwidth values in ► [Table 45.3](#).



■ Fig. 45.11
Freeway overcrossing

■ Table 45.2
Bandwidth per vehicle versus range for 8-lane freeway overpass

Range	Number of vehicles	Channel data bandwidth (Mbps)			
		1	3	10	50
		Bandwidth per vehicle (Kbps)			
<50	64	15.625	46.875	156.250	781.250
75	144	6.944	20.833	69.444	347.222
100	224	4.464	13.393	44.643	223.214
125	304	3.289	9.868	32.895	164.474
150	384	2.604	7.813	26.042	130.208
175	464	2.155	6.466	21.552	107.759
200	544	1.838	5.515	18.382	91.912
225	624	1.603	4.808	16.026	80.128
250	704	1.420	4.261	14.205	71.023
275	784	1.276	3.827	12.755	63.776
300	864	1.157	3.472	11.574	57.870
325	944	1.059	3.178	10.593	52.966
350	1,024	0.977	2.930	9.766	48.828
375	1,104	0.906	2.717	9.058	45.290
400	1,184	0.845	2.534	8.446	42.230
425	1,264	0.791	2.373	7.911	39.557
450	1,344	0.744	2.232	7.440	37.202
475	1,424	0.702	2.107	7.022	35.112
500	1,504	0.665	1.995	6.649	33.245
1,000	3,104	0.322	0.966	3.222	16.108
5,000	15,904	0.063	0.189	0.629	3.144

The ranges used in this table represent the ranges typically provided for various wireless systems. For example, WiFi (802.11 a/b/g) typically have a range of about 50 m or less. Dedicated Short Range Communication/ Wireless Access in Vehicular Environments (DSRC/WAVE) (802.11p) has a usable range of about 500 m, and 3G cellular systems have ranges between 1,000 and 5,000 m.

This table illustrates why DSRC is the medium of choice for vehicle-to-vehicle communication since it provides about 10 Mbps. The effective bandwidth of the system is between 6 and 156 kbps per vehicle. More importantly, the message repeat interval is between 26 ms and 0.6 s. In contrast a 3G cellular system (notwithstanding the addressing issues discussed above) provides about 3 Mbps over a 1,000–5,000 m range. Because the longer range means that more vehicles must share the channel, and the resulting bandwidth per vehicle is between 200 and 1,000 bits/s. In this case the interval until a vehicle can repeat its message is between 4 and 21 s.

■ Table 45.3

Message repeat interval per vehicle versus range for 8-lane freeway overpass

Range	Number of vehicles	Channel data bandwidth (Mbps)			
		1	3	10	50
		Minimum message repeat interval per (seconds/vehicle)			
<50	64	0.256	0.085	0.026	0.005
75	144	0.576	0.192	0.058	0.012
100	224	0.896	0.299	0.090	0.018
125	304	1.216	0.405	0.122	0.024
150	384	1.536	0.512	0.154	0.031
175	464	1.856	0.619	0.186	0.037
200	544	2.176	0.725	0.218	0.044
225	624	2.496	0.832	0.250	0.050
250	704	2.816	0.939	0.282	0.056
275	784	3.136	1.045	0.314	0.063
300	864	3.456	1.152	0.346	0.069
325	944	3.776	1.259	0.378	0.076
350	1,024	4.096	1.365	0.410	0.082
375	1,104	4.416	1.472	0.442	0.088
400	1,184	4.736	1.579	0.474	0.095
425	1,264	5.056	1.685	0.506	0.101
450	1,344	5.376	1.792	0.538	0.108
475	1,424	5.696	1.899	0.570	0.114
500	1,504	6.016	2.005	0.602	0.120
1,000	3,104	12.416	4.139	1.242	0.248
5,000	15,904	63.616	21.205	6.362	1.272

The key observation for this discussion is that the density of the roadway environment argues for smaller RF footprints so that the channel bandwidth is distributed over a smaller population, and thus each user has more bandwidth.

6 DSRC Overview

DSRC provides data communications between terminals located in relatively close proximity to each other. Ranges are typically less than 100–200 m, although in some environments reliable communication has been observed up to about 800 m.

DSRC supports two different types of messaging: broadcast and unicast. Broadcast messages are sent with no network or destination terminal address. They are intended to be received by all terminals in radio range. Unicast messages are addressed to one target

terminal. Since, in a radio environment, all terminals receive all transmitted messages, Unicast messages are filtered at the receiver. If the message address does not match the address of the receiver, then the message is simply discarded.

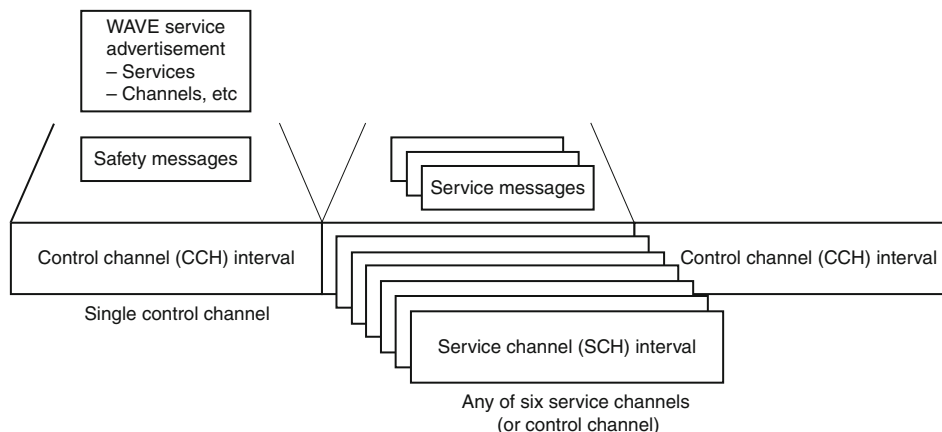
Because of the short range nature of the system, DSRC is uniquely effective for delivering broadcast messages that relate to the local area of transmission. This local characteristic is especially useful since, in general, the traveling public desires to remain anonymous for privacy reasons, and the local nature of DSRC allows communication of location relevant information without any need for conventional network addressing. Most conventional data communications systems rely on some form of node address to target messages to specific terminals. However, in the roadway environment, the mobile terminals are generally in motion, and it is not feasible (or desirable for privacy reasons) to keep track of the location of each addressed terminal. Instead, the system can simply broadcast information in a small region limited by the range of the RF system, and any DSRC terminal in range can receive it. This is particularly effective for mobile to mobile applications where it would not make any sense to try to communicate with another vehicle several blocks away, or for road hazard applications where the mobile terminals that need to receive the information are only those traveling on the affected road. By contrast, using an addressed system to deliver such a roadway specific message would require first determining the addresses of all of the vehicles that should receive the messages, and this is impractical, especially in a large-scale environment where literally hundreds of millions of mobile nodes may be present, and there may be millions of messages, each one relevant only to a very small number of terminals (based on their location).

6.1 Communicating Entities

The DSRC system defines two types of communicating entities: Providers and Users. The Provider advertises services by broadcasting a special message known as a WAVE Service Advertisement (WSA). This message includes a number, known as a Provider Service Identifier (PSID), for each type of service it provides. In general, providers are fixed terminals located along the roadway, although there is no restriction preventing a mobile provider. The WSA includes a list of PSIDs corresponding to the services available from the provider along with the channel(s) that those services will use. The WSA allows users to then determine if the Provider is offering any services of interest to its applications. It is important to note that both Providers and Users may send messages at any time. The only difference is that the User does not broadcast a WSA.

6.2 Channels

The current DSRC standards divide up the 75 MHz spectrum into 10 MHz channels. This allows fixed terminals (known as Roadside Units, or RSUs) in local proximity of each other to provide services without causing interference.



■ Fig. 45.12
DSRC channel management concept

Since it is critical for safety reasons to ensure that all terminals can hear each other, and the standards developers did not want to assume the use of multiple radio receiver systems (or very wide band receiver systems), a method for channel management is described in IEEE 1609.3 and IEEE 1609.4. The approach splits the use of channels into two time intervals (of 50 ms each) called the Control Channel (CCH) interval and the Service Channel (SCH) interval as shown in Fig. 45.12. All terminals are required to monitor the CCH during the CCH interval. In Provider mode, the terminal transmits a WAVE Service Advertisement (WSA) on the CCH during the CCH interval, and since all terminals are monitoring this channel at that time, they all receive the WSA. The WSA contains a list of the services that that Provider will support during the SCH interval along with the SCH channel number they will be using. If a terminal in user mode receives a WSA that contains a service of interest, the terminal will switch to the appropriate SCH during the SCH interval, and will make use of that service. The mechanism for describing these services is described below.

Because all terminals are required to monitor the CCH during the CCH interval, all high priority safety messages are sent on the CCH during the CCH interval.

All low priority services are restricted to only use the SCH during the SCH interval. The result of this method is that all terminals have a high probability of receiving important messages, and less important message traffic is distributed across the other channels, thereby reducing congestion.

7 Mobile Terminals


Since DSRC is ideally suited to the roadway communications environment, it is useful to examine the types of terminals that this system might use. Other communications

systems, such as cellular, are possible as well, but these systems are more familiar to readers, so they will not be discussed here.

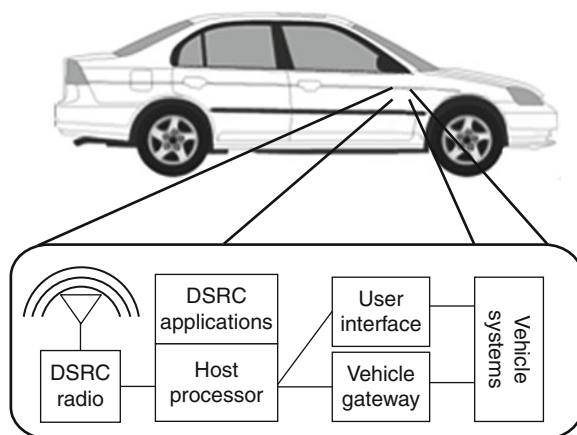
A mobile terminal is a generic term describing a device that is capable of exchanging information with its peer terminals or with fixed infrastructure terminals using a wireless medium. The device may be moving or stationary when operating. Various types of wireless media may be used, although, because the mobile terminals are often moving at relatively high speed through a complex environment, some wireless systems are more appropriate than others. The wireless devices may be embedded in vehicle original equipment, part of equipment retrofitted into a vehicle possibly as an aftermarket product, or part of a portable consumer electronic device. These different terminal implementations are described in the following sections.

In most implementations the mobile terminal typically includes interfaces to the user and to local equipment, and, since generally vehicle communications is location based, these terminals typically include or have access to positioning estimation, for example the Global Positioning System (GPS).

7.1 Embedded Vehicle Terminal

A typical embedded vehicle terminal is shown in  Fig. 45.13. This implementation includes an interface that allows the collection of a variety of vehicle data that can then be sent over the DSRC link. Depending on the implementation, this interface may be a bidirectional gateway allowing authorized input of data to the vehicle, or it may be a one-way data reporting gateway.

The DSRC radio is typically supported by a host processor that runs various applications that use the DSRC system. In many embedded systems the user interface will be



■ Fig. 45.13

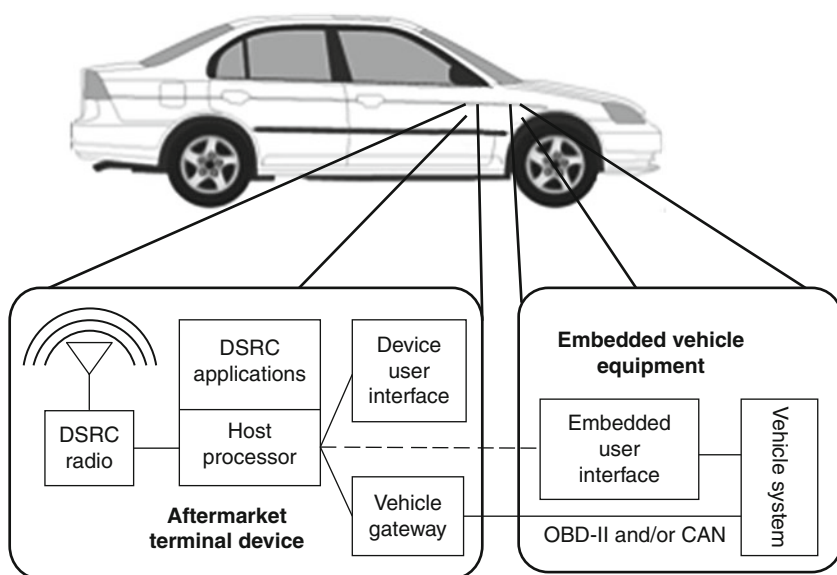
Embedded DSRC terminal structure

implemented as an integral element of the vehicle, and the host processor and vehicle gateway will be deeply embedded as well. In general, embedded vehicle implementations are exclusively controlled by the vehicle Original Equipment Manufacturer (OEM).

7.2 Aftermarket Vehicle Terminal

Aftermarket vehicle terminals are closely related to embedded terminals except they depend on an aftermarket physical installation in the vehicle, and will typically include their own dedicated user interface. Depending on the origin of the terminal, the vehicle interface may include extensive vehicle data (for example if the aftermarket device is approved by the vehicle manufacturer) or it may be limited to data available through the mandated On-Board Diagnostics (OBD-II) connector.

A typical aftermarket implementation is illustrated in [Fig. 45.14](#). In addition to variations in the vehicle interface, it is expected that some advanced implementations may also take advantage of specialized user interface technologies such as Ford Sync[®]. These systems allow third party devices to access a user interface provided by the manufacturer embedded in the vehicle. This approach is attractive since it assures a high quality user interface that complies with vehicle manufacturer safety objectives but does not depend on the long vehicle product development cycle, so it can support a changing variety of aftermarket terminal implementations.



■ Fig. 45.14

Aftermarket terminal structure

7.3 Portable Consumer Electronic Terminal

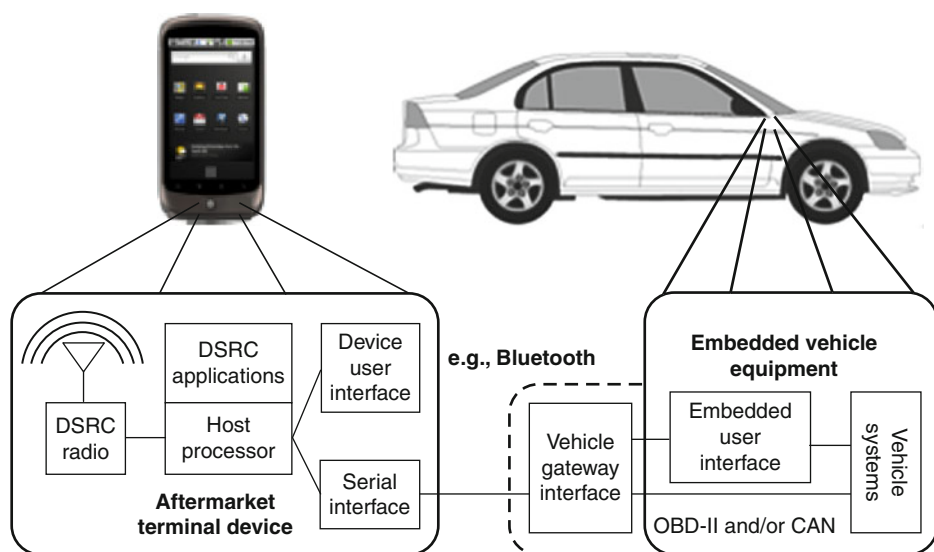
Consumer terminals may take the form of DSRC functionality embedded in common portable electronic devices. For example, [Fig. 45.15](#) shows a conventional smart phone with embedded DSRC capability. This device may interface, for example, using Bluetooth, with a vehicle interface, to acquire vehicle data. Using this approach, the consumer could take advantage of a variety of DSRC based services outside the vehicle, and then they could access higher-level vehicle related services while in the vehicle. In addition, this approach may provide a simple way to rapidly increase the population of equipped users.

7.4 Infrastructure Terminals

A roadside unit (RSU) is a stationary or fixed DSRC transceiver that is positioned along a road. A movable RSU may be mounted on a vehicle, be hand carried, or be a stand-alone device.

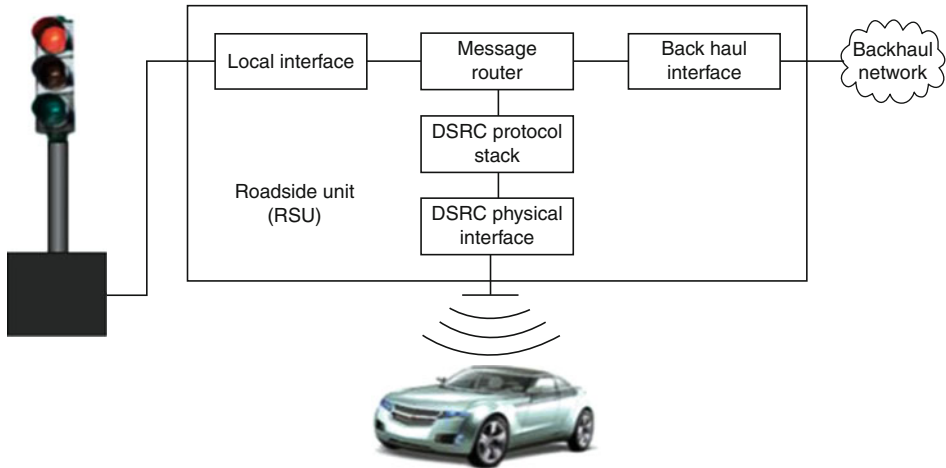
It is important to note that in many applications substantial additional functionality is bundled with the RSU. In this case it is common to refer to the combined unit as Roadside Equipment (RSE). An RSE thus might include all of the RSU functions plus a variety of local services and applications, all operating in a single hardware unit.

The RSU antenna is typically mounted about 5 m above the roadway. The height of the installation can have a significant impact on the overall RF performance as discussed earlier in this chapter. A typical roadside terminal setup is illustrated in [Fig. 45.16](#).



■ Fig. 45.15

Consumer electronic DSRC terminal



■ Fig. 45.16
Typical RSU structure

8 Example Implementations

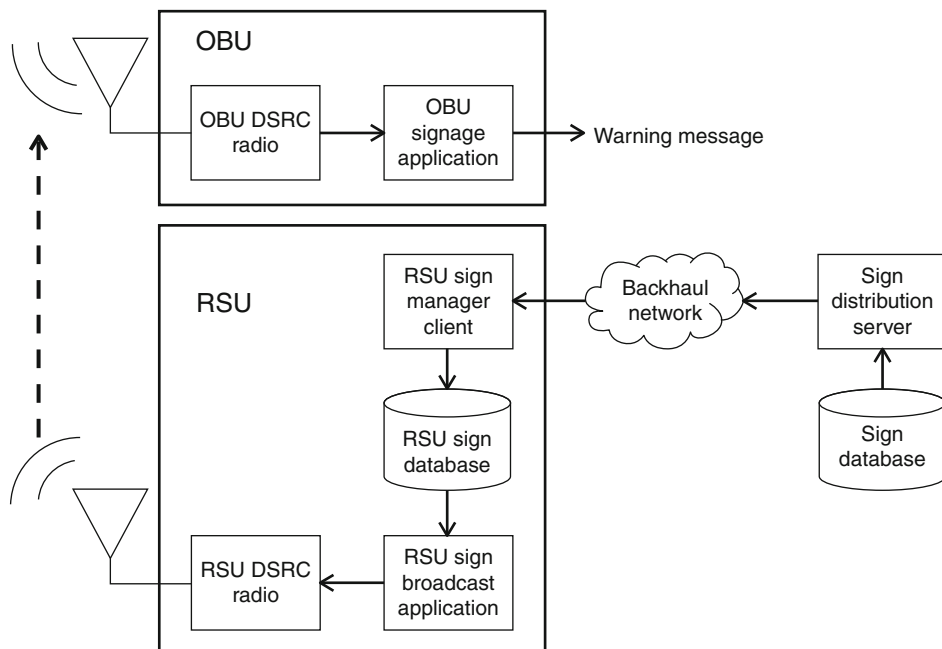
This section describes DSRC operations in the context of several example applications: Signage, Probe Data Collection, Traffic Signal Violation Warning, and Road Hazard Warning. In each case, the way in which DSRC is used for the application is described.

8.1 Signage

Signage can be used to convey information about upcoming features of the roadway environment. This information may be related to important road hazards, in which case it is considered high priority, or it may relate to convenience elements such as roadside services, or general road condition information such as traffic congestion. One set of signage message formats is defined by the SAE J2735 standard.

Typically the RSU is set up with an application that can be accessed over the back haul network from a remote server, so that signs can be “placed” in the RSU. The RSU application then transmits them on the appropriate channel during that channel interval, and repeats the transmission according to the broadcast instructions that accompany the message. An example architecture for this is illustrated in ● Fig. 45.17 below.

In this system a person or system authorized to transmit signs selects a sign from a database of sign types. They may modify variable parameters of the sign to conform it to the particular situation in which it is to be used. They also define the broadcast instructions, which would include the location of the broadcast (e.g., the RSU identifier, or location), the channel for broadcast, the broadcast repeat interval, the times of the broadcast, and the duration of the broadcast. This information would presumably include some security and



■ Fig. 45.17

Typical signage architecture

authorization information as well. The sign distribution server then sends this information over the back haul network to a corresponding client application running on the RSU. This application receives the sign and stores it in a database. It also passes the broadcast instructions to an RSU sign broadcast application. The sign broadcast application establishes a broadcast schedule for all of the signs it is directed to broadcast. At the appropriate intervals, it then submits the signs to the DSRC radio, which then broadcasts them on the appropriate DSRC channel. In practice there are many ways to implement this architecture. For example, for very simple systems, the “back haul” network could be a mobile DSRC transceiver, and the sign could be delivered to the RSU over the DSRC channel from a passing maintenance vehicle. In very simple cases the sign could be hardwired into the RSU, although this would severely limit the flexibility of the implementation.

The DSRC system provides different delivery mechanisms for high priority and low priority information, as described below.

8.2 High Priority Signage

High priority sign messages are broadcast on the CCH during the CCH interval. In the figure above, the RSU sign broadcast application would submit the sign message to the DSRC radio and designate that this was a CCH message. The radio would then transmit the message during the CCH interval.

8.3 Lower Priority Signage

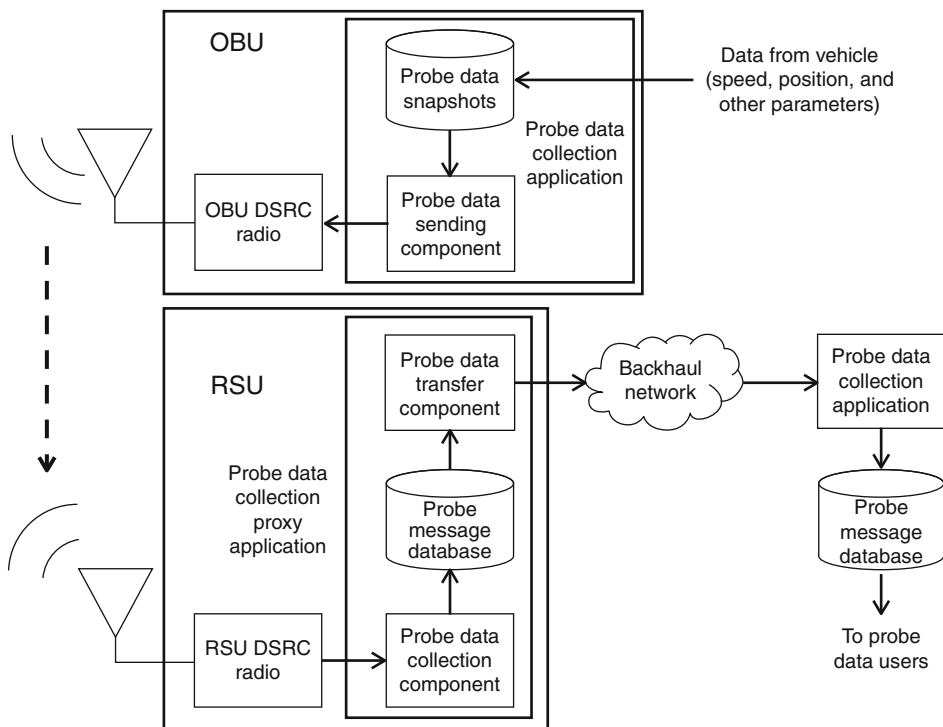
Lower priority sign messages must be broadcast on a service channel. This is slightly more involved than the CCH case above. In this case the RSU sign broadcast application must register the sign service with the DSRC radio. This will cause the PSID for the service to be broadcast in the RSU's WSA. The WSA also provides a field that the RSU Sign Broadcast application can optionally use. This is called the Provider Service Context (PSC) field. The application may use this to provide additional information relating to the sign or the service. For example, the application might include a serial number for the sign set it is currently providing. If the signs have not changed, the serial number would be the same. When the OBU DSRC radio receives the WSA, it notifies the OBU signage application that a registered service is available. It also provides the PSC associated with that service. The application can then examine the PSC and determine if the service is relevant. For example, if the serial number of the sign set has not changed, there is no need to visit the service channel and re-receive the same signs. If the serial number has changed, or if it is new, then the application notifies the DSRC radio that it wanted to access the service, and during the SCH interval the radio would receive the broadcast signs.

8.4 Probe Data Collection

Probe data collection is a process used to collect vehicle operating information and send it to a central repository. Depending on the architecture, the data may be stored, it may be processed in some way, or it may be distributed to subscribers in real time.

The probe data process is generally described in the SAE J2735 standard which describes the message formats, and sets forth rules to preserve privacy. There is, however, no requirement in DSRC to use this standard, and other approaches could be implemented. A conceptual probe data collection architecture is provided in [Fig. 45.18](#).

In this figure, data is collected by the Probe Data collection application on the OBU. This data might include speed, position, time, and other vehicle data such as headlamp status, wiper status, etc. This data is typically gathered at intervals and stored as what are known as "snapshots"; these snapshots basically capture the vehicle state at regular intervals. Because the OBU is not always in the vicinity of an RSU, the snapshots are saved in a memory. When the OBU encounters an RSU that is offering the probe data collection service, The OBU DSRC radio will receive a WSA containing the Probe Data Collection PSID. It will then notify the probe data collection application that the service is available. The probe data collection application will compile a probe data message that includes the snapshots stored in the memory. As noted above, SAE J2735 imposes a variety of rules on how the snapshots are managed to mitigate various privacy concerns. When the probe message is compiled, it is provided to the radio which transmits it on the channel identified in the WSA. The RSU DSRC radio receives this message, and passes it to the probe data collection proxy application. This application resides on the RSU, and essentially transfers received probe data messages back to the central probe data collection



■ Fig. 45.18

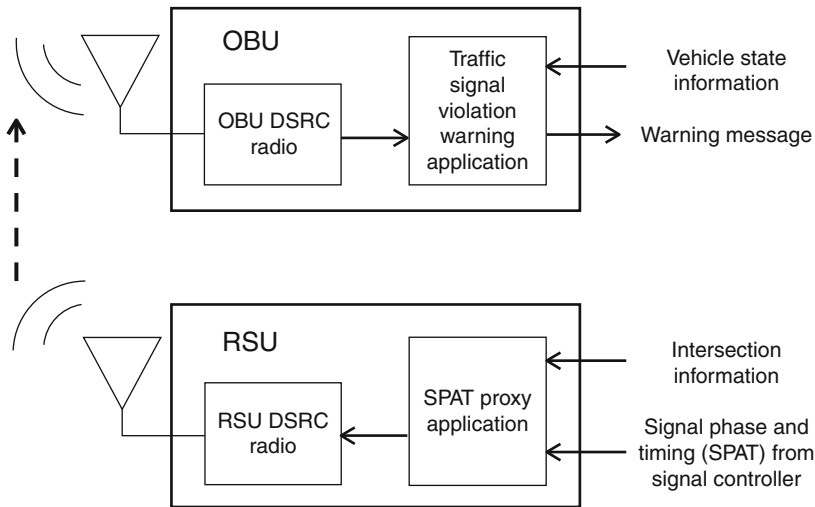
Probe data collection architecture

facility. This is used because there are typically many RSUs collecting data, and it would require too much backhaul bandwidth to distribute probe data from each RSU directly to each data user (not to mention very complex to manage such an arrangement).

The Probe Data Collection Proxy may gather data from multiple vehicles, and pass this data in large blocks to the probe data collection system, or it may pass each message individually. This depends on the particular implementation. When the probe data message(s) is received, the probe data collection proxy application sends it over the back haul to the central probe data collection application. This application is presumably receiving a large volume of data from many RSUs. This data is then either stored for analysis, or, depending on the implementation, it may be sent directly to users.

8.5 Traffic Signal Violation Warning

From a DSRC perspective, the traffic signal violation warning application is similar to the signage application. The difference is in the way the message is created, and how it is used.



■ Fig. 45.19
Traffic signal warning components

An example traffic signal violation warning application is shown in Fig. 45.19. The traffic signal controller provides data about the state of the traffic signal. This typically includes the signal state at a point in time, and the amount of time it will be in that state before it changes to the next state (this is often called Signal Phase and Timing, or SPAT). The SPAT message is then submitted to the RSU DSRC radio for transmission on the CCH during the next CCH interval. Timing is very important in this application because from the moment the SPAT information is provided to the RSU, it is aging. If it is delivered too late, the signal may change state before the message is sent. The SPAT information is usually provided to the RSU through a direct local connection, for example an Ethernet or serial link.

The RSU traffic signal violation warning application may also send information about the intersection. This is typically a geometric description of the locations of the limit lines and the lanes, together with any relevant operational information for each lane (for example, protected turn lanes, etc). This information is typically sent on the SCH and as a result it is advertised in the same way as a lower priority sign.

On receiving the WSA, the OBU traffic signal violation warning application will access the advertised service on the SCH, and will receive the intersection information. It will then receive the SPAT information via the CCH during each successive CCH interval. This application uses the intersection information, together with vehicle data such as speed, position, and acceleration to determine if the vehicle dynamics are appropriate for the signal state. For example, if the state is “green” and it is expected to remain green for a duration long enough for the vehicle to pass through the intersection, it will do nothing. If, on the other hand, the vehicle is proceeding toward a yellow light that will turn red momentarily, and the vehicle is not slowing down sufficiently to stop safely, the application will warn the driver.

9 Conclusion

This chapter has outlined the key challenges and requirements imposed on vehicle communications. These challenges are described in terms of various operational contexts, specifically:

The geographic context, where the different physical scales or ranges involved in communications were discussed, and the types of technologies useful to address communications in those different spheres were described. These were defined as the personal sphere, the region in the immediate close proximity to the user, the local sphere, the region in the vicinity of the vehicle, and the regional sphere, the region in which the road network resides. Each of these scales presents unique challenges to the communications process.

The temporal context, where the moving nature of vehicles was examined in terms of message timing, encounter time, and limitations on the types of network protocols that can be used in this dynamic environment.

The reliability context where a variety of RF propagation phenomena that introduce failure modes in the communications process was examined. These include range effects, multipath, hidden terminal effects, and channel congestion.

Cellular systems provide a good solution for the regional sphere, and wired or serial wireless systems provide a good solution in the personal sphere. DSRC appears to be uniquely suited for the local sphere where the need for communications depends on the location of the vehicle, or the proximity of the vehicle to a place, or to another vehicle. Cellular, Bluetooth, and serial wired systems have been generally covered in other technical publications, so they were not discussed further here. DSRC is not well understood, so the basic system was described, and key elements related to the demands of the local context were examined.

A series of examples were provided to describe different types of mobile terminal and to outline the operation of several local sphere applications that can be implemented using DSRC.

References

- 802.11p-2010: IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments, IEEE 2010
- IEEE trial-use standard for wireless access in vehicular environments (WAVE) – multi-channel operation (IEEE 1609), 2006
- Final report: vehicle infrastructure integration proof of concept technical description, United States Department of Transportation, FHWA – JPO-09-017
- Final report: vehicle infrastructure integration proof of concept results and findings summary – vehicle, United States Department of Transportation, FHWA – JPO-09-043
- Ramo S, Whinnery JR, Van Duzer T (1994) Fields and waves in communications electronics, 3rd edn. Wiley, New York
- Tanenbaum AS, Wetherall D (2011) Computer networks, 5th edn. Prentice Hall, London

46 Vehicle-to-Vehicle (V2V) and Vehicle-to- Infrastructure (V2I) Communications and Cooperative Driving

Scott Andrews

Systems Engineering and Intellectual Property Management in the
Automotive, Mobile Computing, and Communications Domains,
Cogenia Partners, LLC, Petaluma, CA, USA

1	<i>Communications Supported Intelligent Vehicle Systems</i>	1123
2	<i>Vehicle-To-Infrastructure (V2I) Systems</i>	1124
2.1	V2I Applications	1125
2.1.1	Static Events	1125
2.1.2	Dynamic Events	1126
2.1.3	Data Exchange Applications	1127
2.2	V2I Application Physical Setup	1129
3	<i>Vehicle-To-Vehicle (V2V) Systems</i>	1131
3.1	Cooperative Driving Overview	1131
3.1.1	Basic Safety Message	1132
3.2	Longitudinal Trajectory Applications	1132
3.2.1	Cooperative Cruise Control	1133
3.2.2	Cooperative Forward Collision Warning/Avoidance	1133
3.2.3	Emergency Electronic Brake Light	1134
3.2.4	Do-Not-Pass Warning	1134
3.3	Lateral Trajectory Applications	1135
3.3.1	Blind Spot/Lane Change Warning	1135
3.3.2	Highway Merge Assistant	1135
3.4	Crossing Path Applications	1135
3.4.1	Intersection Collision Warning/Avoidance (Through Path)	1136

- 3.4.2 Intersection Collision Warning/Avoidance (Turning Path) 1136
- 3.4.3 Intersection Movement Assist 1136
- 3.5 Specialized Situational Applications 1137
 - 3.5.1 Emergency Vehicle Warning 1137
 - 3.5.2 Slow/Stopped Vehicle Ahead/Work Zone Warning from Maintenance Vehicles/Post-Crash Warning 1137
 - 3.5.3 Pre-crash Sensing 1137
 - 3.5.4 Control Loss Warning 1138
- 3.6 Cooperative Communications 1138
 - 3.6.1 Vehicle-Based Road Condition Warning 1138
 - 3.6.2 Wrong Way Driver Warning 1138
 - 3.6.3 Mayday 1139
- 4 **Communications Requirements** **1139**
 - 4.1 Range 1139
 - 4.2 Link Budget 1141
 - 4.3 Transmit Power and Sensitivity 1141
 - 4.4 Additional Reliability Elements 1142
 - 4.5 Localization Versus Bandwidth and Capacity 1143
 - 4.6 Networking 1143
- 5 **Conclusions** **1144**

Abstract: Cooperative vehicle to infrastructure and vehicle to vehicle safety and mobility applications are described. Applications are separated into static and dynamic categories, such as fixed roadway hazards versus changing traffic signals. Communications requirements are discussed in the context of overall application functional requirements including the specific safe distances required for driver to react to alerts and warnings and to take appropriate action. These requirements are translated into conventional link budget terms relative to those that affect communication reliability such as a transmitter power and receiver sensitivity, multipath fading, and overall channel capacity. The implications of networked transmissions versus broadcast, non-network transmissions are discussed in the context of safety applications.

1 Communications Supported Intelligent Vehicle Systems

Communications based vehicle systems include both Vehicle-to-Infrastructure (V2I) systems and Vehicle-to-Vehicle systems (V2V). V2I systems include communications from the infrastructure to the vehicle and also from the vehicle to the infrastructure. These systems typically support five different classes of application.

Informational applications, sometimes referred to as *signage* applications, simply present an electronic indication to the driver that a hazard or control *may* be present ahead. They are intended to provide an increased level of awareness to the driver as he proceeds along the road. These applications generally provide an electronic in-vehicle display or presentation (possibly including audio) of common road information signs as described more fully below.

Alert applications are based on more specific information that a hazard or control is actually present ahead. Examples of this include end of queue alerts, ice alerts, etc. Alert applications generally require some form of external activation that informs the system that the hazard or control is actually present.

Warning applications typically include electronic alerts and signs that take into account the current state of the vehicle, and, based on the situation, may warn the driver of an imminent threat. Warnings are often based on the vehicle speed and other higher order position related parameters (for example, deceleration, or the lack thereof).

Control applications take the warning process a large step further by assuming some degree of control of the vehicle in order to avoid the hazard or conform to the traffic control. Examples of control applications include speed adaptation, where the vehicle speed is automatically adjusted to fall below the speed limit, automatic braking for signals, etc.

Data Exchange applications involve transferring some specific data between a vehicle and a service provider. These exchanges may be static or dynamic. Static exchanges do not necessarily involve a request or notification. Instead they occur based on some internal trigger, for example, for the collection of probe data, which, depending on the design of the application may simply occur when the vehicle has accumulated a specific volume of data, or when it determines that it is at a location where the exchange can be executed. Dynamic exchanges include an exchange of at least two messages, as in

a request/response scheme. Examples include toll payment transactions, and various freight clearing transactions.

Information, Alert, Warning, and Control applications typically involve a broadcast of information, since the roadway information is generally useful to all vehicles on the road. These applications typically do not involve any sort of response from the receiver of the information. It is possible to also deliver this information on a one-to one basis, although this is generally inefficient. This is discussed on more detail in [Sect. 4.6](#).

All five of these application types may use some combination of exclusive application specific messages, or they use common messages that provide specific information that is used in different ways by different classes of application.

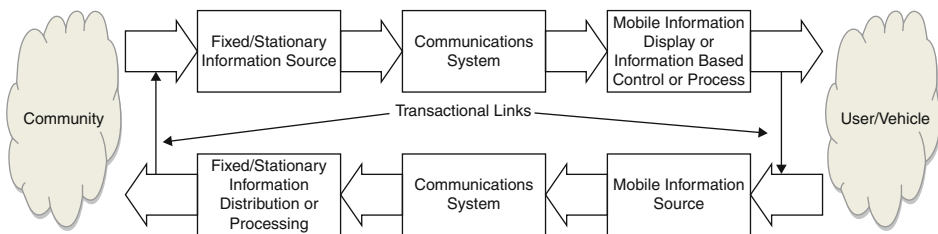
As described in [Sect. 4](#), the geographic nature of these different types of application has a substantial influence on the type of communication system that is best suited to support them.

2 Vehicle-To-Infrastructure (V2I) Systems

Vehicle-to-Infrastructure (V2I) communications supports a wide array of applications aimed at as safety as well as mobility. Safety applications will generally involve alerts and warnings, while mobility applications may involve collecting data from vehicles in order to better understand the overall traffic situation in a region, or it may involve providing wider area situational awareness, for example, traffic information to vehicles. Mobility applications also include a variety of transactional operations such as freight clearing or electronic tolling.

V2I systems generally address a wide range of geographic scales. For example, some systems may provide a location specific warning near a known fixed hazard, while other systems may provide wide area information.

In general, however, the basic arrangement is either that roadway information is generated by one or more sources, and this is communicated to vehicles, or information is generated by vehicles and is communicated back to the system. In a few instances, for example, transactions, an application may involve both of these operations. This is illustrated in simplified form below ([Fig. 46.1](#)).



■ Fig. 46.1
Overall V2I architecture

In general information conveyed by a V2I system will include some form of geographic identifier so that the information can be used in a location-based context. The specific type of information provided determines the form of the communication system that can be used to deliver this information.

2.1 V2I Applications

As described above, V2I applications generally either provide information about the roadway situation so that the vehicle can either alert or warn the driver, or so it can assume some form of control to avoid a hazard or otherwise mitigate a risk, or they execute some type of payment or clearance transaction.

Information, alert, warning, and control applications typically relate to some type of “application event” and often include positioning information related to the location at which the application must decide to act, or not. In general, application events (e.g., hazards) may be static or dynamic.

2.1.1 Static Events

Static events are applicable all the time, or for extended periods of time (e.g., several days). These events are typically also supported by roadside signage that is either permanently or temporarily placed. United States signage is defined in the US DOT Manual on Universal Traffic Control Devices (MUTCD) (2009), and in most cases in-vehicle presentation of the applicable information will use graphics based on the MUTCD prototypes. European countries generally follow the Vienna Convention on Road Signs, first approved in 1968, although European countries generally modify the specific implementation for local needs and, where appropriate, languages.

A typical static event is a road hazard. For these situations a road hazard warning message is generated or initiated by a road authority, typically at a traffic management center. The hazard related message is broadcast to vehicles in the general vicinity of the hazard. Because the hazard message includes the position of the hazard, the receiving vehicle can then determine locally if it represents a threat. For example, the hazard must generally be on the same road being traveled by the receiving vehicle, and it may depend on the direction of travel. If the hazard does represent a threat, the vehicle system may take any of the alert, warning or control actions described above. For example, if the hazard is some distance away, the system may simply alert the driver. This is helpful since the driver is then “primed” to respond when the hazard is actually encountered. This reduces the reaction distance from the decision sight distance to the stopping sight distance (Gulland 2004; Discussion Paper No. 8. A stopping sight distance and decision sight distance 2004). At closer ranges, and typically based on a determination that the driver is not taking appropriate action, for example, changing lanes or slowing

down, the system may then warn the driver, and indicate the appropriate action. In some systems, if the driver fails to respond the system may take automatic action.

Examples of static hazard events include:

- Uncontrolled Cross Streets
- Stop Sign Controlled Intersections
- Dangerous Curve
- Steep Down Hill Slope
- Dangerous, Unsignalized Cross Road Ahead
- Road Ends Ahead
- Roundabout Ahead
- Lane Ends Ahead (Merge)
- Stop Ahead
- Dip in Road Ahead
- Speed Bump Ahead
- Fire Station Ahead: Watch for Emergency Vehicles
- Animal Crossing Ahead
- Falling Rocks Area Ahead
- Trucks Crossing Ahead
- Road May Flood Ahead
- Narrow Bridge Ahead
- Low Underpass Ahead
- Weight Limit on Bridge Ahead
- Bridge Ices Before Road
- School Zone Ahead
- Vehicle/Trailer Height and Width

2.1.2 Dynamic Events


Dynamic events are highly temporal so that the hazard may be present only intermittently. Where static event information is relevant when only received once, dynamic information either needs to be received regularly as the event changes over time, or the event information must be conveyed so that the application can then determine from the current time, and the communicated information what the current state of the event is. Typical dynamic events are:


Variable location dynamic events:

- Road Construction Ahead
- Road Repair Ahead
- Detour Ahead
- Traffic Accident Warning
- Road Closure
- Weather information

Fixed location dynamic events:

- Lane use restrictions by time of day (e.g., carpool lanes)
- Turn restrictions based on time of day or day of week
- Parking restrictions based on day of week (e.g., for street sweeping)
- School Zone based on time of and day of Week
- Animals in crossing ahead
- Icy bridge, etc.
- Signalized Intersection
- Rail Crossing
- Preempt/Priority

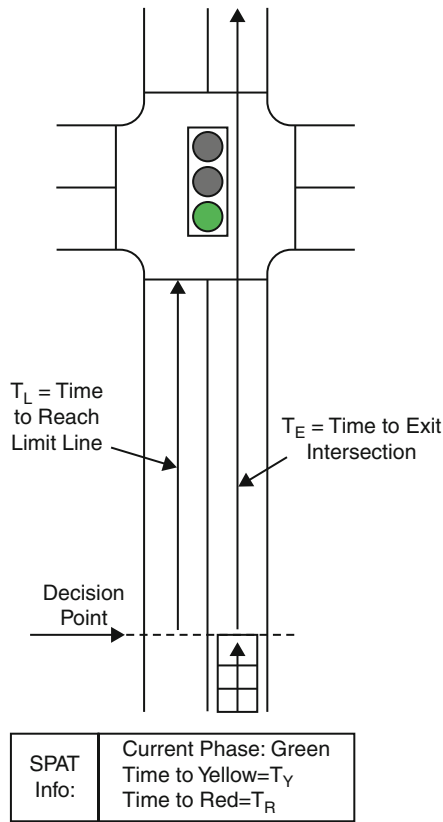
A typical and important type of dynamic event is an intersection. The intersection warning application is similar to the road hazard application except the hazard is the state of an upcoming intersection. Because the signal timing causes the intersection to cycle from being no hazard (green phase), to being a modest hazard (yellow phase), to being a barrier (red phase), the messages must include timing information. Specifically, a Signal Phase and Timing (SPAT) message is generated by the signal controller, based on the current signal state, and is broadcast to nearby vehicles. A vehicle receiving the message can then predict the state of the signal when will be at the entry point of the intersection, and when it will be at the exit point of the intersection. If both of these predictions indicate that the intersection will be safe (i.e., green at entry and yellow or green at exit), the system does nothing. However, if the predictions indicate that the vehicle cannot pass through and exit on at least the yellow phase, the system will warn the driver to stop. As with hazard warning, the system may also include automated braking if the driver fails to respond. This is illustrated in  Fig. 46.2 below.

 Table 46.1 below shows typical scenarios associated with the SPAT application. As can be appreciated from the table, knowing the timing to the next phase and the following phase (subsequent to the next phase) is useful to provide a timely alert if the phase may change while the vehicle is in the intersection.

2.1.3 Data Exchange Applications

Transactional applications typically involve some form of data exchange between a roadside authority and a vehicle. These are typically separated into “clearance applications” and “payment” applications, although other than the end result of the transaction and the types of vehicles involved, there is no significant difference between these, especially in the context of data communications.

As with event based applications, transactional applications may be fixed in location, for example, a toll collection facility at an expressway entrance, or they may be variable in location. This last class represents a relatively new approach wherein road authorities may charge tolls at virtual toll plazas that may change location depending on traffic, day of the week, etc., or they may charge fees based on miles driven, or on current congestion levels.



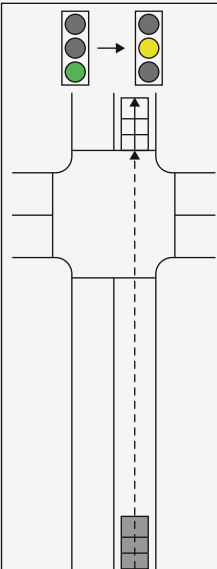
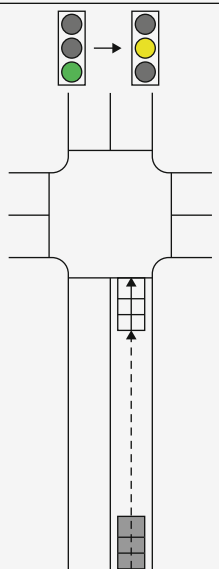
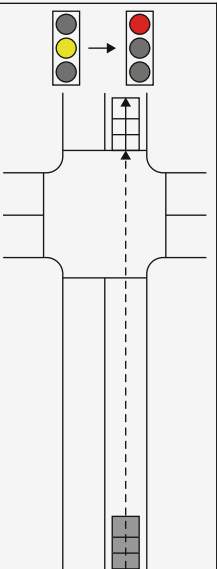
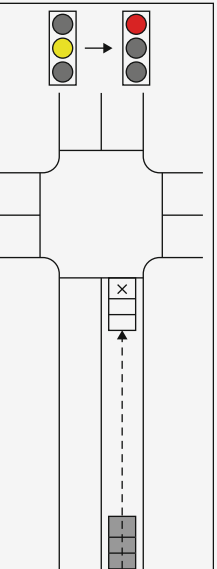
■ Fig. 46.2
Red light running application structure

Typical data exchange events are toll collection and probe data collection. In toll collection, the vehicle is notified of a toll payment region as it approaches the tolling point. This region may be general, or it may be specific to a particular lane, as in High Occupancy Toll (HOT) lanes. As the vehicle enters this region it executes a data exchange with the tolling authority that typically includes payment information. Generally, the transaction is specific to a particular location, and the communications is local, although there is technically no requirement for this. The tolling zones must be small enough to assure that closely spaced vehicles are differentiated from one another.

The probe data exchange is somewhat different. In this application, the vehicle accumulates operational information over some time or distance interval. This information may be simply speed, position, and time, or it may include a variety of other sensor data from the vehicle. This accumulated data is periodically uploaded to a probe data collection system, where it can be processed and redistributed to provide road network state information. This state information may be used by traffic management authorities, or it may be provided to vehicles on the road in the form of real time traffic data. Probe data is described in detail in a companion chapter.

■ Table 46.1

Typical RLR application scenarios

			
$T_E < T_Y$	$T_E < T_Y$	$T_E < T_Y$	$T_E < T_Y$
No Warning	Caution	Extreme Caution	Stop!
Signal will turn yellow after/as vehicle exits intersection	Signal will turn yellow while vehicle is in intersection	Signal will turn red while vehicle is in intersection	Signal will turn red before vehicle enters intersection

2.2 V2I Application Physical Setup

Information, alert, warning, and control applications generally require that the core data relating to the application be conveyed prior to the vehicle reaching an application decision point. The application decision point is the location at which the application must decide to take action. Since the steps of informing, alerting, warning, and controlling typically occur in that order, the decision points for these steps typically occur at decreasing distances from the actual application event. For example, an application might inform the driver that there is a stop ahead at a point relatively far from the intersection. It might then alert the driver, possibly with a higher level of urgency as the intersection nears. It might then warn the driver to stop at a point still closer. This warning point would be located at what is known as the Stopping Sight Distance (Gulland 2004; Discussion Paper No. 8. A stopping sight distance and decision sight distance 2004), which is the point that is sufficiently far from the event (in this case the intersection limit line) that

the driver can perceive the warning, react to it, and then brake to a stop at some nominal deceleration level. The control decision point occurs still closer to the application event, in this case, at the point where the vehicle can still effectively stop itself at a maximal deceleration level (presumably based on an understanding of the current road surface conditions). These different application decision points are illustrated in [Fig. 46.3](#) below.

Data Exchange applications behave somewhat differently. If the communications system is local, such as with DSRC, the application cannot begin the exchange until the vehicle is close enough to assure reliable communications, and the exchange must be completed before the vehicle leaves the radio coverage region. If the communications system is wide area, then the exchange can essentially be executed at any time. Some exchange applications are tied to local enforcement systems, and so it is generally most convenient to also use a local communications scheme. For example, tolling, lane pricing, and freight clearance are generally associated with a data exchange that begins when the vehicle is located at a particular point (e.g., in a toll zone, or at a weigh in motion station). These situations are illustrated in [Figs. 46.3](#) and [46.4](#) below.

[Figure 46.4](#) illustrates that the exchange is limited to the region in which reliable communications can be implemented. This depends primarily on the transmitted signal levels, the sensitivity of the receivers, and multipath effects. Obviously, the duration of

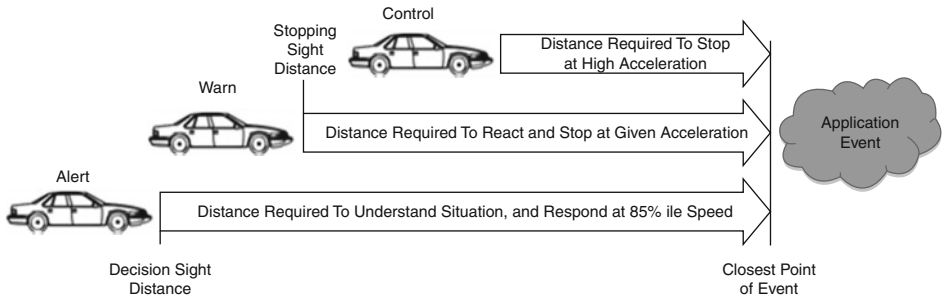


Fig. 46.3
Distances associated with decision point applications

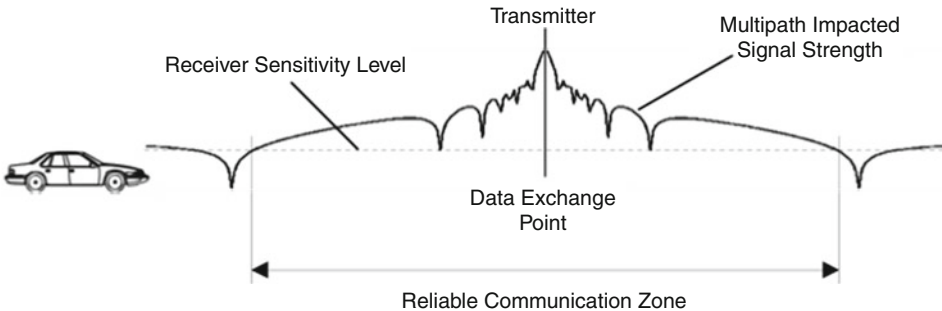
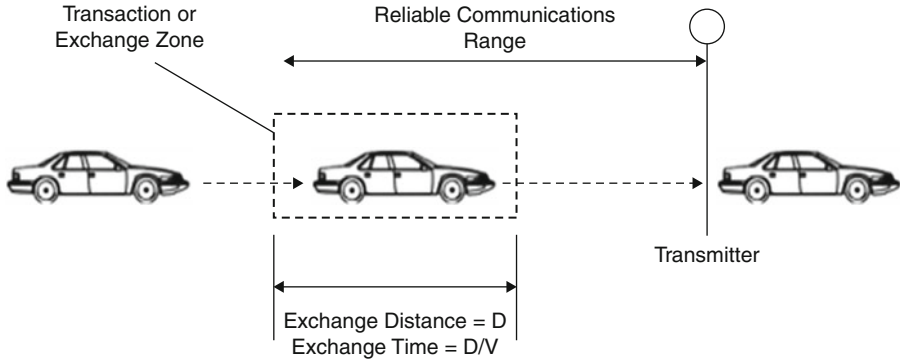


Fig. 46.4
Local data exchange bounded by reception range



■ Fig. 46.5

Local data exchange bounded by defined physical zones

time that the vehicle remains inside the reliable communications zone combined with the bandwidth of the communications system determines the amount of data that can be reliably exchanged.

► *Figure 46.5* illustrates a more constrained situation wherein the exchange must take place within a much more limited geographic zone. These zones are often closely related to various types of enforcement systems, or other sorts of sensing systems. For example, the transaction/exchange zone identified in the figure could be the field of view of an enforcement camera, so that if a vehicle enters the field of view of the camera and fails to execute a transaction, the system will photographically record its license plate number. Alternatively, the exchange zone might be coincident with a truck scale, so that there is a way to bind the data exchanged to the weight of the truck currently on the scale. Because these zones may be quite small to eliminate ambiguity, for example, if there are multiple vehicles present, the vehicle may pass through them very quickly, so the data exchange must also be very short and/or operating at high data bandwidths.

3 Vehicle-To-Vehicle (V2V) Systems

3.1 Cooperative Driving Overview

Cooperative driving applications involve the transmission of messages from one vehicle to the other (generally) nearby vehicles. Conceptually, the vehicles can improve situational awareness by sharing situational information directly. The core message for this application is the Basic safety message, described below.

There is a wide range of opinion about the required range of these systems, and, while the basic concept does result in improved situational awareness, it is inherently unreliable, since it depends on external factors outside the control or knowledge of the receiving or user vehicle. Specifically, most V2V applications are based on determining the state of

other nearby vehicles through the communication of one or more messages. However, if a nearby vehicle fails to send this message, either because its system is inoperative, or because it is not equipped to do so, it is effectively invisible to the system. Since such a vehicle may still represent a physical hazard or threat, the safety system effectively fails to sense this hazard. Since the safety system failure rate requirements are typically less than 1×10^{-2} (SIL 1 or higher), the overall system cannot achieve this failure rate until 99% of vehicles on the road are equipped with the system, and operating properly. This level of fleet penetration and fleet reliability may not be realistically achievable. The counter argument to this point is generally that the driver has a responsibility to pilot the vehicle, and so the system is simply a backup. For example, if a vehicle's brake lights are out, the driver is still responsible for observing that the vehicle is slowing down. However, this is not exactly the same situation, since the driver would be responsible for also observing the brake lights. In these applications the system is responsible for the observation, and in many cases it will fail to do so. Nonetheless, there is a great deal of interest in these systems, and research is proceeding apace, and in situations where both vehicles are equipped, the system has been shown to provide effective results.

3.1.1 Basic Safety Message

The core V2V element is the Basic Safety message, or BSM. The BSM is a message defined by the Society of Automotive Engineers (SAE) J2735 standard. That provides a mechanism for communicating a variety of vehicle state information, including position, heading, speed, etc. It is unclear if the current content of the Basic Safety Message is sufficient to support all V2V applications, but the intent is that this message should be able to be used by many different V2V applications, and few, if any, additional application specific messages will be necessary. The SAE J2735 message standard is undergoing a review sponsored by the US DOT, and it remains to be seen if the existing messages will continue as they are defined, or if they will be modified to support a higher level of state information, and consequently will support a broader array of applications. It is assumed for many of the applications described below that the BSM provides the critical information needed to properly inform the host application.

3.2 Longitudinal Trajectory Applications

Longitudinal trajectory applications deal with the forward trajectory of the vehicle relative to other vehicles. This may be influenced by the vehicle speed or by steering, either of which will alter the space and time trajectory of the vehicle. Longitudinal warning applications alert or warn the driver so that they can control the vehicle appropriately, while control applications use braking, throttle, and/or steering to change the longitudinal trajectory. Longitudinal applications include cooperative cruise control, forward collision warning, forward collision avoidance, emergency electronic brake light, and do-not-pass warnings.

3.2.1 Cooperative Cruise Control

Adaptive cruise control systems adjust the host vehicle speed to match the lead vehicle speed when the lead vehicle is traveling slower than the set cruise speed. Typically these systems are implemented using a forward ranging sensor such as radar or lidar. However, because the roadway environment is highly complex, it is often challenging to separate actual vehicles from other objects when measured by a simple ranging device. Communications based cooperative cruise control allows the implementation of a cruise control system with either no forward range sensor, or with a substantially simpler sensor. Essentially the system uses information from the BSM to determine the trajectories of vehicles ahead. Using the position, speed, and heading of the host vehicle, it determines the host vehicle trajectory. If the system includes ranging sensors, it can then use this trajectory information to select specific range returns that correspond to vehicles in the lane of travel ahead, and to ignore other range returns that are not in the lane of travel. The ranging sensor is then used to determine precise distance to the lead vehicle and the vehicle throttle and brakes are controlled to maintain a safe following distance. In systems that have no ranging sensor, the distance to the vehicle is determined by comparing the reported position of the lead vehicle to the position of the host vehicle. Obviously, in either system lateral positioning accuracy and accurate digital maps are essential since the application must decide if a lead vehicle is traveling in the same lane as the host vehicle. In sensor-less systems the longitudinal accuracy and noise characteristics of the positioning system are critical since noise in position will be observed as noise in the distance, and this may disrupt the control algorithms. In these systems time synchronization is also critical since the positions of the vehicles may or may not correspond to the same instant in time.

In general the communication range requirement for cooperative cruise control systems is about 100 m.

3.2.2 Cooperative Forward Collision Warning/Avoidance

The Cooperative Forward Collision Warning application is designed to alert the driver if there is a risk of a frontal collision (i.e., a rear end collision with a vehicle in the lane ahead, or a head on collision with an oncoming vehicle). The same basic application can also be used to control the vehicle to avoid a collision, for example, by applying the brakes to avoid a rear end collision. The application receives basic safety messages transmitted by surrounding vehicles, and determines their trajectories. Using the position, speed, and heading of the host vehicle, it determines the host vehicle trajectory. It then identifies any vehicles traveling in the same lane. It continuously compares the host vehicle trajectory with the trajectories of the other in-lane vehicles, and if a collision may occur, it either warns the driver or controls the vehicle speed, or does both successively.

Obviously this application requires sufficient trajectory accuracy to determine if the other vehicle represents a threat, and in the case of oncoming vehicles (i.e., vehicles

approaching in the same lane as the host vehicle) the range must be sufficient to provide sufficient time for the driver to react to the warning. Since these are unexpected events, the reaction time must be assumed to correspond to the decision sight distance (2.5 s). For the worst case head on collision situation at a 200 kph closing speed, 2.5 s represents 67.5 m, so the range of the communication system must be at least this, plus whatever distance/time is required to actually react. At a lateral acceleration of 0.3 G, a full lane change requires 1.6 s which consumes an additional 44 m, so the total communications range requirement is about 110 m.

3.2.3 Emergency Electronic Brake Light

The Emergency Electronic Brake Light (EEBL) application is used to warn other vehicles behind if a vehicle is stopping quickly. EEBL typically uses a special message to convey this state information, although this type of state information may be introduced into the BSM at some point. Upon heavy braking, the EEBL host vehicle broadcasts an EEBL message that is then received by surrounding vehicles. An EEBL equipped vehicle receiving the message then determines the relevance of the event (e.g., in lane, ahead, behind) based on the trajectories of the vehicles as described above, and, depending on the threat level, may provide a warning to the driver. This application may be particularly useful when the driver's line-of-sight is obstructed by other vehicles or bad weather conditions (e.g., fog, heavy rain), although wireless communication may also be disrupted in these situations as well. The EEBL application requires a range of at least the decision stopping distance. At 100 kph, this is about 150 m.

3.2.4 Do-Not-Pass Warning

The Do-Not-Pass Warning (DNPW) application uses the speed and positions of surrounding vehicles to determine if the host vehicle can safely pass slower vehicles in the same direction of travel without danger from oncoming vehicles. This is a further refinement of the forward collision warning application, except the application must determine the distance required to change lanes, speed up to a passing speed to clear the vehicle(s) in the direction of travel, and then change lanes back into the original lane of travel. If this time is less than the time required for an oncoming vehicle to reach the location where the host vehicle will return to the original lane of travel (obviously with some safety margin), the system will indicate that it is not safe to pass. Such a system either must be active all the time, or it requires some form of activation when the driver intends to pass a vehicle ahead. At 100 kph, assuming the host vehicle accelerates to 20 kph faster than the vehicle being passed, and assuming lateral and longitudinal acceleration levels of about 0.3 g, an entire passing maneuver requires about 8.5 s. At a closing speed of 61 m/s (220 kph, assuming the passing vehicle has accelerated to 20 kph faster than the vehicle being passed), the distance required to pass is about 514 m. This distance will be longer by about 200 m for each additional vehicle being passed.

3.3 Lateral Trajectory Applications

Lateral trajectory applications deal with the lateral trajectory of the vehicle relative to other nearby vehicles, typically in adjacent lanes. These applications typically receive basic safety messages (BSM) from other nearby vehicles, and compare the trajectories of vehicles in adjacent lanes with the host vehicle trajectory. If these trajectories are coincident, the application will warn or control the vehicle. The lateral position is obviously influenced by the vehicle steering, although to some degree the vehicle speed may impact the timing of lateral trajectories. As with longitudinal applications, lateral warning applications alert or warn the driver so that they can control the vehicle appropriately, while control applications use braking, throttle, and/or steering to change the lateral trajectory. Lateral applications include Blind Spot and lane change warning, highway merge assistant and other similar applications.

3.3.1 Blind Spot/Lane Change Warning

The blind spot/lane change warning application provides situational awareness to the driver relative to vehicles that are nearby in immediately adjacent lanes. The application may provide this information in an advisory manner, so, for example, the presence of a vehicle in the blind spot may be indicated, or it may provide a warning based on either the activation of a turn signal when an adjacent vehicle is present or from the determination that the trajectories of the adjacent vehicle and the host vehicle are coincident. Generally the blind spot and lane change applications require very short communication ranges since the vehicles are immediately adjacent.

3.3.2 Highway Merge Assistant

The highway merge assistant application is a combination of lane change warning, and rearward collision warning/avoidance. As with many other applications, this application receives BSMs from nearby vehicles; based on the vehicle position and a digital map the application can determine that it is in a merging situation. Using the trajectories of vehicles approaching from the rear, and the trajectory of the host vehicle, the application can provide guidance on when to accelerate and how hard to accelerate to effect an on-ramp merge. It may also (presumably should) indicate that it is not safe to start the merge when there is insufficient gap to the approaching vehicles.

3.4 Crossing Path Applications

Crossing path V2V applications typically use the BSM to determine the trajectories of vehicles on an intersecting or crossing roadways. SAE J2735 also includes an intersection

collision avoidance (ICA) message that describes the trajectory of the vehicle in the context of the intersection. As with other V2V applications, if the vehicle trajectories are coincident, then the systems will warn the driver, or control the vehicle to avoid the hazard.

Because other V2I systems may exist to warn drivers if they are about to violate a traffic signal or stop sign, crossing path applications are generally considered secondary to the V2V solution. They may be useful in situations where there is no signal control (e.g., blind intersections), but they are not the primary method to avoid a crossing path accident.

3.4.1 Intersection Collision Warning/Avoidance (Through Path)

The intersection collision warning/avoidance application determines the trajectories of all V2V equipped vehicles in the vicinity of the intersection. For the through path situation, if any of these trajectories are coincident with the through path trajectory of the host vehicle, the system will either warn the driver or it may automatically respond, for example, by applying the brakes. These applications are somewhat problematic since vehicle trajectories may change quickly. For example, a vehicle waiting to make a left turn has a static trajectory, so, it represents no threat until it moves and begins to turn, yet at this stage it may be too late to avoid a collision. Vehicles approaching from the side may or may not have an unobstructed communications path, and so they may be unobservable until both vehicles are close to the intersection, at which there also may not be time to react.

3.4.2 Intersection Collision Warning/Avoidance (Turning Path)

As with the through path situation the intersection collision warning/avoidance application determines the trajectories of all V2V equipped vehicles in the vicinity of the intersection. Based on driver activation of the turn signal, the application can predict its changing (future) trajectory. If any of the other vehicle trajectories are coincident with the projected trajectory of the host vehicle the system (i.e., through the turn) the application will either warn the driver or it may automatically respond, for example, by applying the brakes. In the case of a situation where the host vehicle stops to wait for a clearing in traffic before turning, the application can indicate to the driver if it is unsafe to make the turn (because there is insufficient gap to make the turn and accelerate to road speed).

3.4.3 Intersection Movement Assist

Another application relating to crossing path situations is the intersection movement assist. In normal driving, this application would perform the same function as either case of intersection collision warning/avoidance described above. In a more futuristic implementation such an application could be used to modulate the flow of vehicles through

a signal-less intersection. Used in this way the application would cooperate with similar applications in other vehicles. Vehicles in the same or opposing direction of flow would align themselves together, and synchronize their motion with similarly aligned groups of vehicle in the crossing path direction. At modest traffic volumes it may be possible to implement free flow intersections where the vehicles never stop, but instead simply interleave themselves at the intersection. Obviously such an approach requires very high levels of reliability as well as an assurance that 100% of the vehicles are equipped and operating properly.

3.5 Specialized Situational Applications

There are a number of other V2V applications that do not fall into the longitudinal, lateral, and crossing path geometries. These typically involve unusually roadway situations and typically involve specialized messages as opposed to the BSM.

3.5.1 Emergency Vehicle Warning

The Emergency Vehicle Warning application uses a specialized emergency vehicle Alert (EVA) message to notify vehicles in the vicinity that it is in an emergency state. There is little additional information such as trajectory or direction of travel, although this could be provided in an emergency vehicle BSM.

3.5.2 Slow/Stopped Vehicle Ahead/Work Zone Warning from Maintenance Vehicles/Post-Crash Warning

This is a class of application that relates to informing approaching vehicles of an unusual situation. For example, a vehicle stopped by the side of the road can broadcast a message indicating that it is stopped and where it is stopped. The type of vehicle and the reason for the stopping may alter the type of message, for example, a maintenance vehicle might issue a work zone warning message, while a disabled vehicle might issue a beacon that warns approaching vehicles and also notifies a roadside service vehicle as it approaches. Generally the range requirements for these applications are the same as for corresponding V2I (MUTCD) alerts and warnings.

3.5.3 Pre-crash Sensing

The pre-crash sensing application is a mechanism whereby vehicles that are on a collision course that cannot be avoided, provide information about themselves to the other vehicle.

This information can be used to control and preset various crash safety systems to mitigate the effect of the crash. Because vehicles in the final stages of a collision trajectory are assumed to be relatively close, the range requirements for this application are very limited, but the channel access priority requirement may be very high.

3.5.4 Control Loss Warning

The control loss warning application enables an equipped vehicle to broadcast a self-generated, control loss event to surrounding equipped vehicles. Upon receiving a control loss event message, an equipped vehicle can compare the trajectory of the sending vehicle with its own trajectory, and determine if the loss of control represents a threat. If so, the application can alert the driver, or take some other form of avoidance action. Examples of control loss are skidding on ice, approaching a turn too quickly, braking too late at an intersection, etc. As with other local warning applications the range requirements are not challenging, but the latency and channel access requirements are.

3.6 Cooperative Communications

Cooperative communications represents a special class of V2V application that uses other vehicles to relay information down the road by receiving the message and then rebroadcasting it to other vehicles. In general the protocols for such relayed messages have not been highly developed, so these applications are somewhat speculative at this writing. For instance, there is no defined mechanism for physically bounding the extent of the message, so relayed messages could conceivably be relayed over a very large area, and for a longtime span.

3.6.1 Vehicle-Based Road Condition Warning

The vehicle-based road condition warning application involves vehicles sensing road conditions using one or more sensing mechanisms, for example, potholes using accelerometers, or ice using traction control sensors, and then generating and transmitting warning messages providing the nature and location of these hazards. These messages are broadcast to other vehicles and relayed down the road some distance. In this way, vehicles approaching the hazard will receive warning information about the hazard, even though there is no roadside infrastructure to transmit the message.

3.6.2 Wrong Way Driver Warning

The wrong way driver warning application is similar to the vehicle-based road condition warning application, except the hazard is a vehicle traveling the wrong way on the road. The

sensing of this hazard may either be done by the driver, who would then activate the systems, or by the vehicle, based on a comparison of the BSM data transmitted by the other vehicle and the digital map. The relay transmission method is the same as described above.

3.6.3 Mayday

The “Mayday” application provides either manually initiated or automatically initiated emergency situation message to be transmitted following an accident or if the vehicle is disabled. The message is then relayed as described above until it reaches a roadside unit, whereupon it is forwarded to the roadside response authorities. This application would require some mechanism to terminate the relay of the message once it had reached the response center.

4 Communications Requirements

Communications requirements can be quite complex and generally lower level requirements tend to relate to the specific operation of the chosen communications technology. At the application level, communications requirements generally include simply the required range at which the information must be conveyed, the amount of data that must be sent, together with any timing requirements, and the required reliability of the communication. Each application also has particular physical characteristics that may impose topological constraints or demands on the communications system.

From these high level requirements, one can determine the required channel bandwidth, the overall user capacity requirements, the range and physical characteristics of the wireless medium, etc. When a specific technology is chosen, the characteristics of that technology will typically introduce additional operational and performance requirements.

4.1 Range

Vehicles often travel at relatively high speeds, and human drivers have relatively slow reaction times. The decision for an application to take action generally requires determining a point at which the action should be taken such that known events can occur before the actual safety event is reached. For example, if the safety event is a fixed obstacle in the road, and the objective of the application is to warn the driver, then the application must issue the warning sufficiently in advance that the driver has time to perceive the warning, react and then control the vehicle, for example, bring it to a stop.

The AASHTO Green Book (American Association of State Highway and Transportation Officials (AASHTO) 2004) describes two human factors related metrics: the Stopping Sight Distance (SSD) and the Decision Sight Distance (DSD). The stopping sight distance is the distance required to perceive the hazard assuming that the user was expecting it.

■ Table 46.2
SSD (in meters) for various speeds and situations

Speed (kph)	Nominal	Emergency (Wet)	Emergency (Dry)
40	50	27.7	14.2
60	85	59.6	40.3
100	185	163.4	93.4
120	250	235.7	127.9

■ Table 46.3
DSD (in meters) for various maneuvers and speeds

Maneuver	Speed (kph)		
	50	100	120
Stop on rural road	70	200	265
Stop on urban road	155	370	470
Speed/path/direction change rural road	145	315	360
Speed/path/direction change suburban road	170	355	415
Speed/path/direction change urban road	195	400	470

Stopping sight distance is applied where only one obstacle must be seen in the roadway and dealt The decision sight distance includes an added distance associated with the time that the user needs to perceive and understand the hazard. In general if the driver has already been alerted to a potential hazard, then they will perceive it more rapidly that if they happen upon it with no advance preset alertness. Decision sight distance is applied where numerous conflicts, pedestrians, various vehicle types, design features, complex control, intense land use, and topographic conditions must be addressed by the driver.

Generally, one can assume that the DSD should be used as an alert distance, and the SSD can be used as the warning distance.

The perception-reaction time for a driver is often broken down into the four components that are assumed to make up the perception reaction time. These are referred to as the PIEV time or process.

- Perception the time to see or discern an object or event
- Intellection the time to understand the implications of the object’s presence or event
- Emotion the time to decide how to react
- Volition the time to initiate the action, for example, the time to engage the brakes

The AASHTO Green Book provides values for SSD under nominal and emergency conditions. This is provided in ➤ Table 46.2 below for various assumed vehicle speeds.

The AASHTO Green Book also provides DSD under various types of avoidance maneuver. These are provided in ➤ Table 46.3 below for speeds of 50, 100, and 120 kph.

As can be appreciated from the tables above, alerts must be received about 300–400 m from the hazard, and warnings should be received about 200 m from the hazard. This fixes the basic range requirement for most of the V2I hazards, and many of the V2V hazards.

4.2 Link Budget

The link budget is an accounting of the various signal gains, losses, and noise contributions between the transmission of a signal, and the detection of the signal. It includes the transmitted power, the antenna gains (transmitter and receiver), cable losses, receiver noise contributions, and propagation path losses.

The link budget is given by:

$$P_{RX} = P_{TX} + G_{TX} - L_{TX} - L_{FS} - L_M + G_{RX} - L_{RX}$$

where: P_{RX} = Signal power at the receiver P_{TX} = Signal power at the transmitter
 G_{TX} = Antenna gain at the transmitter L_{TX} = Losses at the transmitter L_{FS} = Free space propagation loss (range dependent) L_M = Miscellaneous losses (including margins for fading, etc.) G_{RX} = Antenna gain at the receiver L_{RX} = Losses at the receiver

The processes for developing the link budget are well documented. The critical element for this discussion is the relationship between the physical and operational needs of the application, and corresponding reliability requirements. For a given level of required communications reliability, the minimum power at the receiver, P_{RX} must exceed the noise at the receiver by some predetermined level. This is the signal to noise ratio (SNR), which depends on the energy per bit and the thermal noise in 1 HZ of bandwidth (E_b/N_o), and the ratio of the system data rate and the system bandwidth (R/B_T). System Bit Error Rate (BER) is a direct measure of communication reliability in the absence of other error correcting mechanisms. Once A BER is defined based on the needs of the application, the E_b/N_o can be determined based on the BER and the system modulation. The receiver sensitivity P_{RX} can then be determined from the required bandwidth and the bandwidth of the channel and the E_b/N_o . Once the various losses and margins are accounted for, the values of P_{RX} and P_{TX} can be determined, and this defines the required characteristics of the transmitter (output power) and receiver (sensitivity) for the communication range required by the application.

Thus determining the required range, and necessary bandwidth and the required communications reliability defines the basic requirements for the communications system.


4.3 Transmit Power and Sensitivity

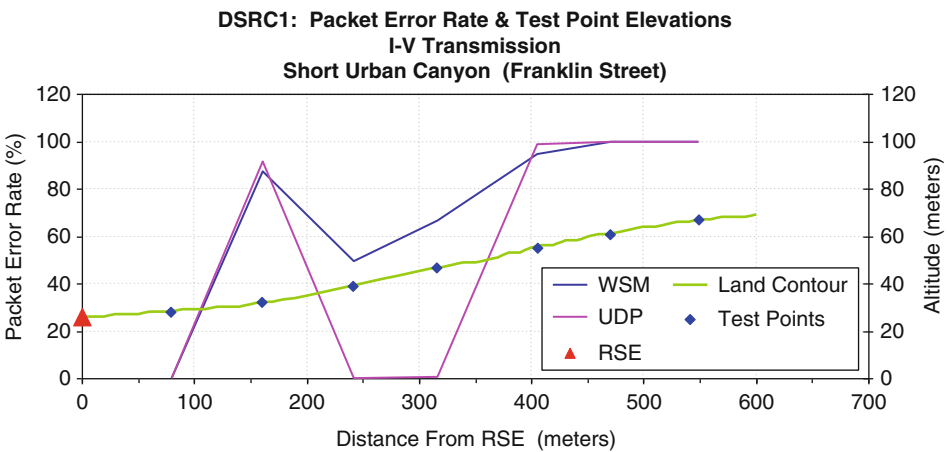
Since transmit power and receiver sensitivity are key to achieving the required communication reliability at the specified bandwidth, it is necessary to specify these in a true operational sense. Specifically, the gain pattern of an antenna mounted on a vehicle may be quite different from that same antenna mounted on a ground plan in a test chamber.


Similarly, roof slope and height, or other body features on the vehicle may alter the on-road behavior of the system. This means that receiver sensitivity and transmitted power (EIRP) may be characterized against required values in relation to the vehicle axes with the vehicle at rest on a flat road. The gain patterns must also account for variations in longitudinal body tilt under acceleration and deceleration loads, and must allow sufficient margin to account for road junctions that may be at different slopes.

Transmit power and receiver sensitivity also need to account (in the link budget process) for multipath fading. Most recent tests indicate the need for some form of multipath fading mitigation (e.g., diversity antennas).

4.4 Additional Reliability Elements

In some situations it may not be possible to achieve the required reliability level with single message transactions. In these cases it may be desirable to repeat messages some number of times. This effectively adds redundancy to the communication system, and reduces the overall message failure rate at the expense of additional data throughput demand.  **Figure 46.6**, below, shows the difference in packet error rate in a DSRC test carried out during the VII Proof of concept test (United States Department of Transportation 2009a, b). This figure shows packet error rate as a function of distance for single packet messages send using the WAVE short message protocol (IEEE 1069.3 WSM (IEEE 2006)), and the conventional UDP/IP protocol. The WSM protocol is broadcast only, and so there are no message acknowledgments, so only a single packet is sent. The UDP protocol is an addressed message, and the system includes automatic MAC layer acknowledgments. With UDP, if the MAC layer acknowledgment is not received, the system will retry until



 **Fig. 46.6**
Message reliability comparison with and without message repeats (Permission from the VII Consortium)

it is received, or until it has tried 10 times. As can be seen in the figure. As the vehicle enters the multipath fading zone between 100 and 250 m, both UDP and WSM transmissions experience substantial packet errors. However, between 250 and 350 m the UDP transmissions improve substantially, while the WSM transmissions only improve slightly. In this case, the message retries in UDP are adding redundancy, and this is improving overall message reliability.

4.5 Localization Versus Bandwidth and Capacity

Generally the applications described in this chapter related to localized communications. In a few cases wide area communications system maybe required, but generally the messages being communicated are local in nature. A wide area communication system, for example, a wide area broadcast system such as HD radio may be useful for this, depending on the radio footprint. While wide area systems are generally cheaper because one hardware installation can serve a large number of vehicles, the larger area also means there may be substantially more messages, because the larger footprint is likely to encompass more events. Wide area two-way communications systems are especially problematic since the number of vehicles technically increases as the square of the range. This result in channel access and capacity issues that range from simply too many vehicles trying to send messages, to very complex hidden terminal issues. Smaller radio footprints limit the number of vehicles seeking to access the channel, and substantially reduce the number of message in the channel, thereby reducing overall channel capacity and bandwidth requirements. Of course the radio footprint must be sufficiently large to meet the range requirements for the application, as described above.

4.6 Networking

Networking is generally not necessary, and in many cases is not desirable for the applications described above. Networking assumes that each member of the network has established a network identity (typically a network address), and that that identity has been communicated to the other members of the network. In this way, any member of the network may send a message to any other member. Some networks, for example, (most cellular systems) are configured so that a base node (base station) can communicate with a large number of mobile nodes, but the mobile notes cannot directly communicate with each other. This approach obviously is impractical for V2V applications. It is also not particularly applicable to V2I applications involving the distribution of alert and warning information. The reason for this is simply that the same information may need to be sent to multiple vehicles in the same area. Using a networked approach, this requires transmitting the same message multiple times. Of course, some networking systems provide for limited broadcast or multicast, which may mitigate this problem.

In general, however, unless the applications are very wide area, For example, covering a metro region, network based systems require substantial management overhead as vehicles enter and exit the network, and since few of the expected applications require one-to-one addressed messaging, this overhead serves no useful purpose. For situation where networking provides value, for example, when executing TCP/IP or UDP/IP transactions, IPv6 is the preferred approach since the stateless auto-configuration process defined for IPv6 allows user terminals to acquire an IP address with no overhead. In these applications, the system is using networking functions, but there is technically no complete network that is being managed (there is no ability, for example, to send an IP packet to another vehicle without first learning the IP address of that vehicle).

5 Conclusions

This chapter has outlined a variety of V2I and V2V applications. In general these applications involve communication of information between the infrastructure and the vehicle or between vehicles over relatively short ranges (ranges that are meaningful in the physical safety context of the vehicle). We have shown the relationships between these physical aspects and the communications requirements, and have illustrated briefly how one can derive communications requirements from the basic application needs of range, bandwidth, and reliability.

References

- American Association of State Highway and Transportation Officials (AASHTO) (2004) A policy on geometric design of highways and streets (The Green Book), 5th Edn, Association of State Highway Transportation Officials. http://downloads.transportation.org/aashto_catalog.pdf
- (2004) Discussion Paper No. 8. A stopping sight distance and decision sight distance, prepared for Oregon Department of Transportation, Salem, Oregon by the Kiewit Center for Infrastructure and Transportation, Oregon State University Corvallis, Oregon Sept 2004
- Gulland WC (2004) Methods of determining safety integrity level (SIL) Requirements – Pros and Cons, April 2004
- IEEE (2006) Trial-use standard for wireless access in vehicular environments (WAVE) – Multi-channel operation (IEEE 1609.3)
- United States Department of Transportation (2009) Final Report: Vehicle infrastructure integration proof of concept results and findings summary –vehicle, United States Department of Transportation, FHWA – JPO-09-043
- United States Department of Transportation (2009) Final Report: vehicle infrastructure integration proof of concept technical description, United States Department of Transportation, FHWA – JPO-09-017
- US DOT Federal Highway Administration (2009) Manual on uniform traffic control devices, 2009 edn, US DOT Federal Highway Administration

47 Probes and Intelligent Vehicles

Christopher Wilson

TomTom Group, San Francisco Bay Area, CA, USA

1	<i>Introduction</i>	1146
1.1	What Are Probes?	1147
1.2	Technology	1148
1.2.1	Vehicle Installation	1149
1.2.2	Communications	1150
2	<i>Applications</i>	1153
2.1	Real-Time Applications	1153
2.1.1	Traffic Reporting	1153
2.1.2	Traffic Smoothing	1157
2.1.3	Active Signal Management	1158
2.1.4	Weather	1159
2.2	Historic Data	1159
2.2.1	Transportation Network Design and Maintenance	1160
2.2.2	Traffic Control Attribute Detection	1161
2.3	Probe Maps	1163
2.3.1	Road Classification	1166
2.3.2	Intersection Geometry	1166
2.3.3	Instantaneous Speed Data	1167
2.3.4	Application Development and Validation	1169
2.4	Privacy	1169
2.4.1	Increasing Probe Penetration	1170
3	<i>Conclusion</i>	1171

Abstract: Positioning and communications technologies are enabling the collection of massive amounts of probe data from the vehicle fleet. The quality and quantity of the data vary significantly depending on technology and collection interval, with the best data coming from GPS systems integrated into vehicles. Real-time data collection provides, for the first time, detailed insight into vehicle behavior throughout the major elements of the transportation network. This is used by roadway managers to optimize performance of the infrastructure and by intelligent vehicles to increase situation awareness beyond the range of their autonomous sensors, potentially leading to significant increases in safety and efficiency.

Historical probe data can be used to create a map of driver behavior at every point in the transportation system. Behavioral maps share some attributes with traditional physical maps, but other attributes, such as average vehicle speed and distribution of speeds, enable novel implementations of intelligent vehicle applications. Behavioral data can be used directly to define normal or acceptable behaviors. In addition, a particular driver's preferred location on behavior distribution curves can be used to predict future behavior and personalize interactions.

Probe data will become more prevalent and valuable as penetration increases and latencies decrease. Ultimately, the need for probe data will help motivate deployment of short-range, high data rate communications between vehicles and with infrastructure.

1 Introduction

Vehicle positioning and affordable communications capability have been available to consumers for about 15 years. This is fundamentally changing the way the transportation system operates since it is now possible to close the feedback loops between vehicles and actuators (generally signals). It is possible to observe in real time with fine granularity such attributes as speed, delays, queue lengths and travel time, and to provide this data, or derivatives thereof, to traffic management centers and vehicles. This data, when widely available, will allow the transportation network to be run as a system with predicted improvements of 30% or more. It also provides intelligent vehicles critical information about the vehicle and infrastructure environment beyond the range of their sensors, and about historical behaviors.

For the first time investigators have large-scale data on that most enigmatic actuator, the human driver. This data provides historical models of driving on every road and through every maneuver which can be segmented by any population identified in the probe data. The average behaviors defined in this way can provide specific feedback to drivers on safety or efficiency performance with respect to other drivers on the same road. Eventually these behavioral models will feed into the algorithms used for autonomous driving, to make autonomous vehicles understand human driver behaviors and to make autonomous actions more human and more comfortable to human passengers.

Data from vehicles on the roads is the only way to provide widespread information on what is happening. All other techniques are either local (e.g., infrastructure sensors) or

based on time-consuming surveys. The communications models are still being developed, along with the business models, and the necessary commitments to privacy, but the future vision of an integrated transportation system is based on probe data.

1.1 What Are Probes?

Transportation probes are vehicles that opportunistically collect data as they traverse the transportation network. They are not driven for the purpose of collecting data and cannot be tasked (although data collection parameters may change) and there is little or no interaction with the driver or other humans in the vehicle. In general, the data is anonymous. From an application point of view, probe data does not provide data on a particular vehicle (e.g., that car that is going to hit me, the car parked in this spot, the first car in the queue for that light), rather the data is representative of an area (e.g., an anonymous car less than 1 min ahead, a car parking nearby, a car in the queue at the light ahead). This is the primary difference between probe-based applications and Vehicle to Vehicle (V2V) applications where critical communications are between specific vehicles, and there is a relationship between the communicating vehicles.

Probe sensors may be permanently associated with a vehicle, perhaps built in by the vehicle manufacturer, or may be temporarily associated with the vehicle, such as a cell phone, personal navigation device, or toll tag. In all cases probes provide location and time information, other data may also be available. Data from a personal device (e.g., cell phone or tablet) must be examined to determine if the device is in a vehicle or not.

The value of probes is that they are inexpensive and potentially prolific, providing valuable information based more on statistics and high penetration than on the accuracy of any individual measurement. The capital and operational costs of each individual probe are low. Value is often generated by aggregating the data from many probes into a coherent picture of some aspect of the transportation system. Even so, probe collection and communications systems usually piggyback on other primary applications since the probe data on its own does not provide sufficient value to support the collection costs. This is usually a vehicle tracking service or a cell phone data plan. In general, probe vehicles are not extensively modified in order to be probes.

In its simplest form, probe data consists of a location and time, direction of travel and speed. This is generally considered to be adequate for traffic data, and has a minimum of privacy issues since the individual observations are autonomous and need not be linked together, making tracking difficult. Often this sort of data is provided by fleet management companies to probe data integrators after anonymizing the data and stripping off proprietary data fields. Individual fleet management companies do not have sufficient vehicles to compute traffic patterns, so traffic data companies, such as Inrix, aggregate probe data from many fleets. Traffic data companies have probe data available from millions of dedicated vehicles providing relatively continuous data with archives of trillions of observations. Often these vehicles are trucks and delivery vans which spend a disproportionate amount of time on the road, thus further increasing the value of this data for traffic purposes. Several tens of millions of vehicles



■ Fig. 47.1

Probe data captured in Munich over a 40-day period with 1-second spacing of observations. The color of each pixel is defined by the number of observations in that cell. Note that road usage is clear and the behavior at intersections can indicate the type of control present (Image courtesy of TomTom, © 2010)

may provide intermittent information as they request traffic information or navigation updates, and send a single probe data packet to a traffic server. There are also companies that have their own proprietary population of probes, such as Navteq, which has unique access to Nokia phones, and TomTom, which has access to some of their 45 million Personal Navigation Devices (PNDs) sold. ➤ [Figure 47.1](#) shows probe data captured for a 40-day period around Munich. This image provides some idea of the massive amounts of data that are available. Since each point also has a time associated, speeds are easily determined.

1.2 Technology

Just about every vehicle on the road has the equipment onboard to serve as a probe. This varies from an OEM installed connected navigation system to an occasionally connected PND on the windshield, to a smartphone with GPS, or a simple cellular phone in the driver's pocket emitting coarsely tracked GSM signals. These are not all of equal value for probe applications. Combining these various systems, and paying for the

communications costs that may be associated, is one of the key aspects of designing a working probe aggregation system. The architecture depends on the applications that will be supported, and will change with time as more high-quality data becomes available.

Vehicle positioning and communications capabilities are the first-order drivers of probe data utility. Positioning is primarily a function of the level of vehicle integration (assuming a GPS sensor), while communications can take many forms from real-time cellular connections to “sneakernet” involving transfer of a physical device from the vehicle to a networked computer.

There are currently four widespread approaches to probe data collection, identified here according to their level of integration into the vehicle. These are (1) vehicle manufacturer installation, (2) aftermarket connection to the vehicle buses, (3) aftermarket vehicle appliance, and (4) personal mobile devices.

1.2.1 Vehicle Installation

OEM installed systems are the most valuable for probe data collection. The sensors are well understood, including their location with respect to the vehicle and the algorithms used for positioning, which often involve sensors beyond GPS, such as wheel rotation, accelerometers, gyroscopes, or steering wheel angle. Positioning may be to lane level accuracy (~ 1 m). These systems also have the potential to provide significant amounts of data from other sensors on the vehicle, such as temperature, rain, or road surface characteristics (see [Table 47.1](#)). In many cases this data exists on the vehicle bus and merely needs to be collected. In addition, limitations on power consumption, weight, and size are minimal and installation issues are eliminated.

Communications vary with the vehicle. High-end vehicles often have an integrated cell phone capable of real-time connectivity (Mercedes, BMW, GM Onstar), some, such as Ford’s Sync, have connectivity through the user’s phone, while other vehicles have devices that can be removed and data collected via a USB connection (Fiat, Mazda, Renault). Most vehicles, even those with navigation systems, have no connectivity.

The primary limitation with OEM systems is that the vehicle manufacturer must be engaged. While all vehicle manufacturers are working on this topic both individually and in various consortia (e.g., <http://www.vehicle-infrastructure.org/>), the business case for OEM-controlled communications is not clear for most vehicles. However in the long term, it is likely that all vehicles will be connected, greatly increasing the number of high-quality probe sources.

Aftermarket connections to the vehicle buses are a good path to having high-quality vehicle data through a retrofit to any vehicle. Several systems currently in operation connect to a vehicle’s On-Board Diagnostics port (OBD). The port has been mandated in vehicles for many years and provides a standard set of information generally related to emissions, including vehicle speed, as well as optional and proprietary data. Applications such as fleet monitoring and insurance telematics often connect this port to a communications unit (often with other functionality) and make some of the data available off-board, possibly in real time. The cost for the connectivity and installation is still prohibitive but as industries

■ Table 47.1
Partial list of attributes available in many vehicles

System	Data
Kinematic	Location
	Altitude
	Speed
	Acceleration
Engine	Fuel flow
	Energy consumption
	Barometric pressure
Cabin environment	Precipitation (wipers)
	Temperature
Drivetrain/brakes	ABS/ESP activation
	Vertical wheel accelerations
RF environment	Cellular connectivity
	GPS signals
Camera or radar	Location of other vehicles (traffic conditions)

appreciate the value of the probe data and gain experience, this market is growing rapidly. This may become a major source of probe data for managing emissions.

Aftermarket vehicle appliances include devices such as personal navigation devices (PNDs) and toll tags, generally associated with the vehicle. Toll tags use short-range communications (which provides the position) whereas PNDs may be connected via a cellular data link or intermittently connected to the Internet via cable or data card. The advantage of these devices is that they are prevalent and easy to install. The data is generally limited to kinematic data such as position and velocity. While these devices are usually in road vehicles (as opposed to a pedestrian or trains), often with penetration rates in excess of 10%, they may be turned off or in the glove compartment seriously impacting collection rates.

Personal mobile devices (such as cell phones) tend to have very high penetration rates and are usually connected, but they generally do not provide precise, frequent locations. Cellular carriers can provide a general location for any phone that is on; however, this is not sufficiently precise for most probe-based applications. To get a good position, a suitable application must be running, using a GPS and providing the position data to an aggregator. Bluetooth signals can also be tracked if a device is in discovery mode. This is commonly done by a pair of devices on the roadside which can track the travel time of a Bluetooth device between the two locations, generally several hundred meters apart (► [Table 47.2](#)).

1.2.2 Communications

Currently there are three communications systems for probe data in widespread use. By far the most common is the cellular network, but for applications that need widespread

■ Table 47.2

Probe collection technologies and the corresponding location accuracy

Approach	Description	Accuracy	Comments	Local infrastructure	Vehicle penetration
Cell phone signal probes	Based on signals sent by all mobile phones that are turned on	50–1,000 m	Accuracy varies widely depending on cellular infrastructure and modifications to cell towers. May be multiple phones per vehicle	N	Carrier's local market share
Cell phone GPS probes	Based on GPS in phone (tablet)	10 m	Uses considerable power. Suitable application must be running	N	Smartphone penetration X application adoption
Tolling systems	The transponder is dedicated for the collection of vehicle tolls, but detectable anywhere with suitable infrastructure	3–30 m	Low-cost probes, but generally high-cost-roadside units with very specific and limited functionality. Penetration can be quite high in some regions, and there is generally a one-to-one correspondence between vehicles and transponders	Y	0–70% extremely regionally-dependent
GPS probes in vehicle (PND or navigation system)	Data from a GPS unit, provided through some application	2–20 m	Generally requires that some navigation or other location-aware application is running in a device. GPS can be very power-hungry and not suitable for phones on battery power	N	1–10% increasing rapidly with connected navigation deployments
Bluetooth probes	Infrastructure-based identification of Bluetooth units in "discovery" mode	10–50 m	Requires that collection units be set up along the roadside. Good for speed measurements on specific roads	Y	10%

coverage, satellite communications are also used (primarily Qualcomm), and some systems make use of physical (sneakernet) connections where wireless connectivity is absent or prohibitively expensive. Wireless systems (especially satellite) are fairly expensive compared to the value of the raw probe data with the result that, in general, the frequency of data communications is determined by the primary application and not by the probe requirements. Communications can be event-based, and ISO has developed standards to manage communications (ISO/TS 25114), but this approach, by its nature, selects data to support the existing probe-based applications (or the primary data collection application); the rest is lost. As will be seen later, there is significant value in learning the normal behavior of vehicles, but the cost to transmit this data must be very low in order to be cost-effective. In sophisticated systems with both real-time and delayed (low-cost) communications options, the more valuable data may be sent over the more expensive real-time data link, while the majority of data awaits a low-cost connection for relay. TomTom uses this strategy with the result that they have over four trillion probe data points, representing driving on all roads and in all conditions.

For real-time connectivity, cellular will predominate as more and more vehicles are connected to the Internet through embedded cellular systems. The latency of the network will decrease and become more consistent, providing probe-based systems with some of the characteristics of a V2V network. While an infrastructure-based communications scheme will never be able to handle the latency demands of some critical vehicle safety applications (such as collision avoidance), future cellular systems will support less-demanding applications, such as hazard warning, intersection control, and general situation awareness. The CoCarX project used a test LTE cellular network to demonstrate vehicle-to-vehicle communications (via infrastructure) with a latency below 100 ms (ERTICO 2011). This technology will only improve.

Ultimately the spectrum constraints and cost are likely to drive probe data collection to short-range Vehicle to Infrastructure (V2I) communications links overlaying the V2V systems where probe data already exists in the message sets (SAE J2735). It seems quite possible that an early use for the dedicated 5.9 GHz spectrum will be to support local vehicle-infrastructure applications, such as Cooperative Intersection Collision Avoidance, wherein probe data is collected from vehicles to define and validate the maps used in the application. This would also be an excellent opportunity to download other probe data, and to upload data to vehicles at relatively low cost. This incentivizes installation of the infrastructure for traffic management. Dedicated Short Range Communications (DSRC) communications systems, as proposed under several connected vehicle programs, are ideal for the collection of probe data since they allow for intermittent V2I connections to download large volumes of probe data quickly at low cost.

Another issue that must be addressed in discussing probe-based applications is possible selection effects due to the population of probe vehicles. As mentioned above, many current fleets are based predominantly on commercial vehicles. While this has certain advantages in that they are likely to be on the road more often, they will tend to be on major roads (especially commercial long-haul trucks), only working business hours, and they may have unique restrictions, such as truck speed limits, restrictions to certain

lanes (that may have a very different speed than the rest of the traffic), and stopping at weigh stations. The truck speed on a steep grade may be significantly lower than that of light vehicles. Certainly truck behavior around high wind advisories and chain controls are very unlikely to reflect the behavior of light vehicles, and these are exactly the conditions under which traffic information can be the most valuable. Some aggregators use predominantly light vehicles, which leads to other biases. For many applications the probe population is a critical factor in the success of the application, and yet the composition of the probes and certainly the identity of individual probes are usually poorly known.

2 Applications

The emphasis to date for companies collecting probe data is in providing data to support traffic applications in well-established markets. Existing traffic information is primarily on freeways; the new probe technologies are beginning to obtain accurate information on arterials with hopes of pushing to ever lower road classes. Beyond traffic, probe applications are being developed to support many other intelligent vehicle applications from traffic smoothing and smart routing, to giving applications a human feel, and providing validation of driver and application behavior.

Applications for probe data can be divided by their requirement for near-real-time data or not. This is an important distinction since probe data collection schemes requiring the collection device (or memory card) to be physically connected (usually via USB) to the Internet provide very large amounts of data, but the latency being typically on the order of weeks. This data is not useable for real-time applications, but it can be used to validate and tune the real-time models. The increased volume of high-latency data makes certain applications much more effective.

The effectiveness of near-real-time applications is largely dependent on the penetration in the vehicle fleet. Below are some examples of predicted penetration required for various types of applications (🔗 [Table 47.3](#)).

2.1 Real-Time Applications

2.1.1 Traffic Reporting

Traffic is the bane of drivers in many cities around the world, and in many of those cities it is essentially impossible to build the additional road capacity to absorb the ever increasing number of vehicles. The general conclusion today is that, since the road network cannot be expanded, it must be managed better, and this requires real-time traffic information. In the ideal case, real-time traffic information is sufficiently detailed that incidents can be spotted very quickly (within 5 min) and resources deployed to deal with them before they significantly impact the traffic flows. The road network should be managed to divert vehicles away from impacted areas until they recover. This often implies diverting vehicles to minor roads,

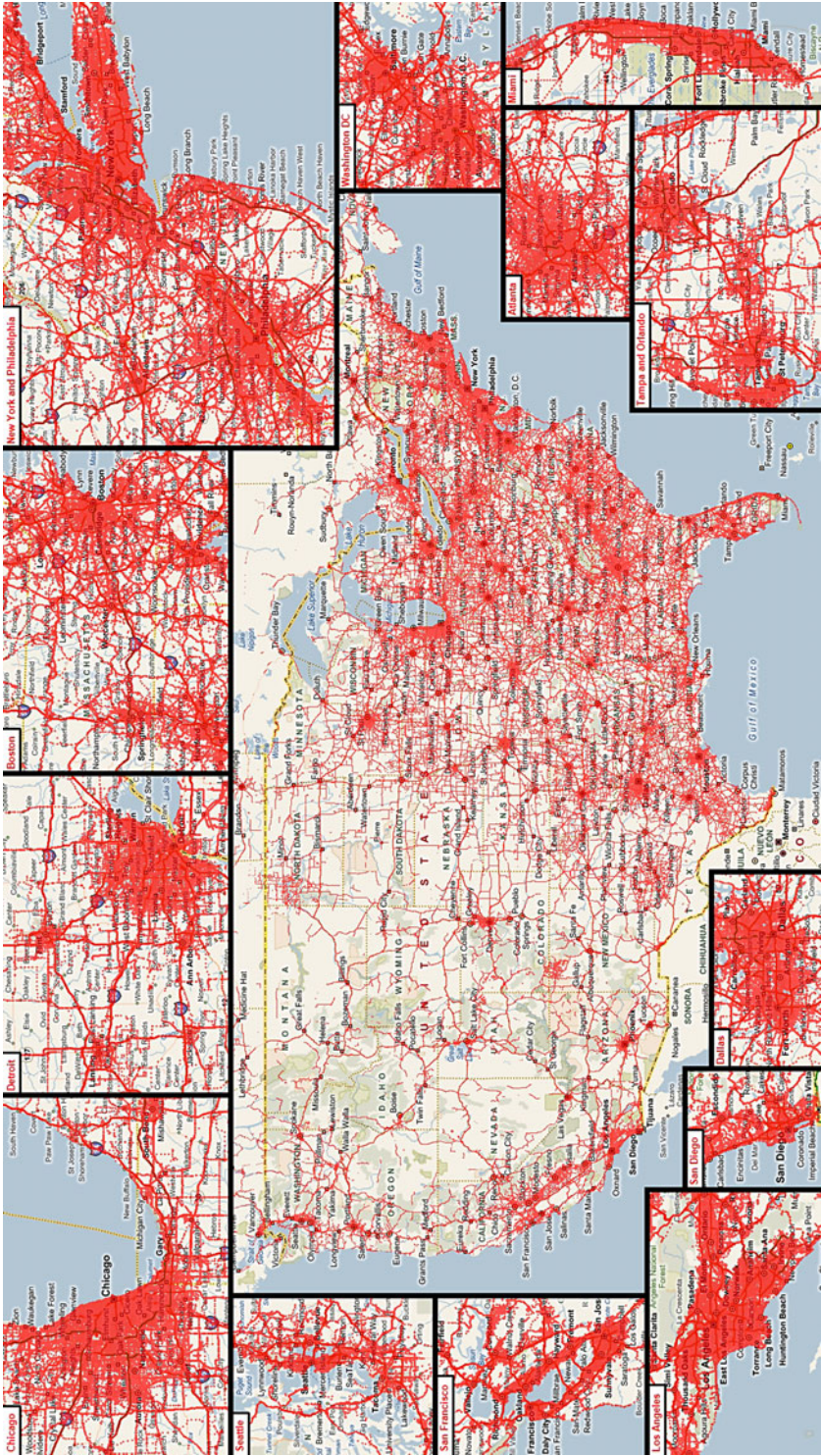
■ Table 47.3
Real-time applications and their typical modeled penetration requirements

Real-time application	Penetration	Description
Traffic on primary roads, “normal” conditions	1%	Identify “familiar” traffic problems on major roads within a few minutes. Extensive use of historical data
Traffic on primary roads	5%	Identify traffic on major roads in unusual conditions. Intelligent Speed Adaptation feasible
Traffic on arterials	15%	Identify congestion, even in the presence of traffic controls and unstable flows. Ability to modify signal timing plans
Actively manage traffic	30%	Modify signal timing. Actively manage groups of vehicles
Infrastructure removed	100%	Infrastructure can become virtual with roadway management done via communications to vehicles from a “dedicated command center”

which requires that the system have information on those secondary routes. Having good traffic information on the minor roads, as well as the major roads, is critical for intelligent vehicles to deal effectively with congestion.

Probe data is the method by which data of sufficient quality can be obtained for all of the critical roads in a region. Traditionally traffic data has been gathered from fixed infrastructure (usually loop detectors cut into the pavement, but also radar and cameras). These systems suffer from several drawbacks, most notable of which is that they only work at specific spots. While not terribly expensive individually, they are very expensive (in both hardware and operational impacts) to install and maintain over a large network, especially if they need to detect incidents quickly which requires spacing on the order of 3–4 sensors per lane mile. In addition, the communications costs can be significant; so even where sensors are installed for local needs, such as managing the queue for a stoplight, these are often not connected into a network. These limitations can be overcome by a probe-based system with sufficient real-time probes to sample the network (Kwon et al. 2007). Most estimates of the probe penetration required for accurate freeway traffic estimation are on the order of 2–5% (this can be lowered significantly by also utilizing historic data, but these models will not work in disaster situations or other times when the historic data is invalid). The penetration required on arterials (where the network is significantly more complex and there are many perturbations, such as signals) is between 5% and 15%. These penetrations have not yet been achieved with GPS probes in real life, and cell phone probes are not sufficiently accurate in a dense road network mixed with pedestrians. ➤ Figure 47.2 shows data collected from Inrix’s fleet of probes during a 15-min interval on a Friday morning. This coverage is generally better than that available from fixed infrastructure, and, unlike infrastructure, probe coverage is expanding rapidly.

The last decade has seen the development of several new ways for collecting traffic data. The deployment of Electronic Toll Collection systems on bridges and roadways has led to



■ Fig. 47.2 Probe data collected by Irix during a 15-minute interval on a Friday morning (Image courtesy of Irix)

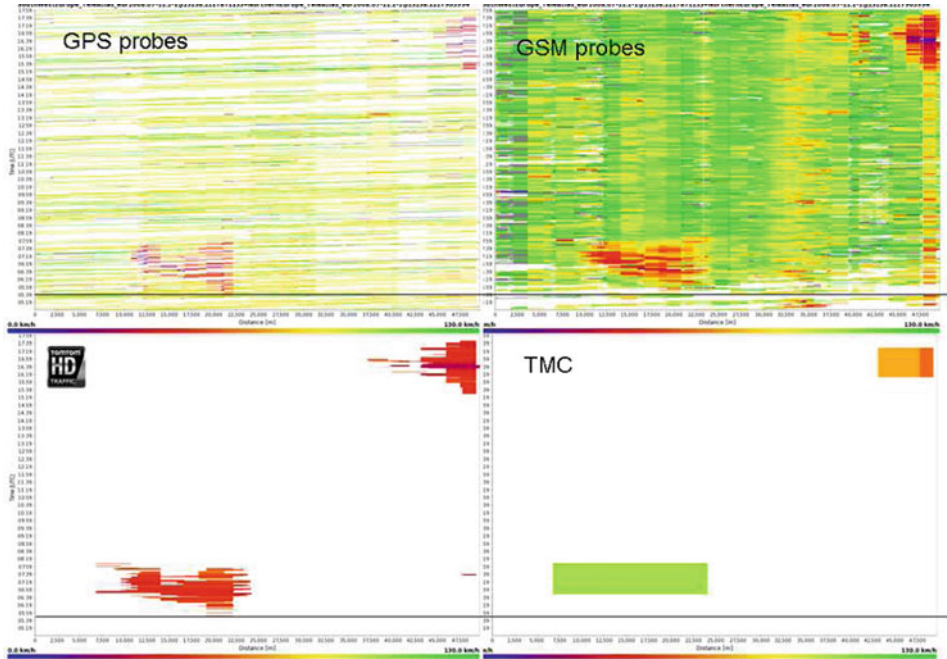
the proliferation of vehicles with short-range transponders that provide a unique identification number to a roadside receiver. Where these tags are prevalent they can act as probes and be tracked with the installation of detectors (which note the passing of a specific tag while not charging any tolls) on signs or other roadside structures. These systems work quite well on major roads in regions where there is a sufficient volume of toll tags to watch, providing accurate travel times between the sensors and definitive information on vehicles' travel paths. These systems are, in general, still too expensive to install on secondary streets (lack of suitable overhead support structures and sheer volume of sensors required). Being intermittent point sensors, they have no way to determine the path traversed between two readers, which is an increasingly severe limitation for lower road classes where the number of probable paths increases quickly. The utility of these tags is limited and they will not be considered further in this discussion.

There have been many attempts, and few successes, in monitoring cell phone signals to determine traffic flows (cell phone probes). Phones are continuously tracked by the phone's service provider in order to route calls to the proper cell, and there are a lot of cell phones to track. By using this location information it is possible to derive the traffic situation. The primary limitation of these systems is the poor location accuracy of individual cell phones, often insufficient to differentiate between adjacent roads based on position alone. In addition, a basic assumption behind cell phone probes is that speed distinguishes phones in cars from phones on pedestrians, (as well as differentiating between phones on a freeway and on the adjacent frontage road). This assumption fails at high congestion levels when all roads are slow, and when the data is most needed.

There are ways to significantly increase the accuracy of cell phone positioning, but these require extensive modifications to the cellular infrastructure, which do not support the carriers' core business of providing communications. In the end, these systems are rapidly being replaced by systems that use data from a GPS embedded in the cell phone or other mobile device. A comparison between cell phone probes and GPS probes can be seen in ► *Fig. 47.3* for the same time period and stretch of road.

The GPS probe category includes many devices, such as smartphones and tablets, as well as navigation systems and vehicle location systems which are installed in the vehicle for other reasons, but have GPS capabilities. These form the basis of all the successful probe-based traffic systems today. The location accuracy of these systems is on the order of 10 m, which is generally good enough to resolve the difference between two adjacent roads.

Smartphones that contain GPS devices and sophisticated applications for working with the data are becoming quite common and are likely to dominate traffic probes in the years to come (until OEM probes take over). In many ways, these are ideal GPS probes – they have GPS, the power to process and work with the data and they are connected. Many smartphones contain one or more navigation applications that can easily be modified to provide probe information. Tests, such as the “Mobile Millennium” (<http://traffic.berkeley.edu/>) show significant progress, especially in addressing privacy concerns, but still suffer from the need to have a particular application running on the cell phone, and the need to have the phone plugged into the car since the GPS is a very significant power



■ Fig. 47.3

Comparison of GPS to Cellular probes. This shows a section of the A3 in Germany with several construction zones. GPS probes show better accuracy and consistency, although fewer probes. The *bottom figures* show the reported traffic using either standard TMC location referencing (based on predefined tables), or TomTom's OpenLR service, which allows referencing arbitrary locations (Image courtesy of TomTom, © 2010)

drain on a phone battery. Most phones will last significantly less than an hour running a GPS-based navigation application on battery power.

The discussion above focuses on real-time traffic, but ultimately, road operators would like to be able to predict traffic, at least for short intervals into the future. Here again the probe data is proving invaluable. Historic probe data can provide a wealth of observational information on how the roadway has behaved in the past under any particular set of conditions, and there is far more historic probe data than there is real-time data. Given the large database of scenarios in the historic record, a relatively few real-time probes can be used to determine which scenario is likely to play out for short-term traffic predictions.

2.1.2 Traffic Smoothing

One application showing significant promise for near-term deployment in intelligent vehicles is traffic smoothing. Traffic smoothing vehicles use speed information from vehicles a short ways ahead to determine the average speed over the next few minutes, and then drive

that speed. This does not affect the total trip time at all, yet it can result in a significant reduction in oscillations in the traffic stream which are often associated with accidents. It can also lead to significant drops in emissions, in some models up to 50%.

Real-time penetration rates in some regions may be sufficient to support traffic smoothing today, and there are a number of evaluations underway. Computer simulations indicate that even with 5–10% probe penetration there could be up to 25% reduction of CO₂ emissions and 50% of other pollutants (NO_x, CO, HC) (Jin 2011) (University of California, Irvine. Personal communication). Researchers are working to determine how to compute the correct average speed based on a relatively small probe sample. A major problem appears to be uncertainty as to how people will react, even if provided with perfect information. On the one hand, this type of system may open large (temporary) gaps immediately in front of the participating vehicles. Adjacent cars may jump into the gap, decreasing the effectiveness, and drivers may not understand the system well enough to leave the gap with the knowledge that they will get the time back when they hit the slow traffic ahead. There is a large education component needed to deploy these systems, and the data to convince users that the system actually works. The existence of Automatic Cruise Control (ACC) may make these systems more acceptable as the traffic smoothing system may input a set speed to the cruise control system. It remains to be seen if the behavior of adjacent cars will make this a comfortable experience for the driver.

2.1.3 Active Signal Management

The key to real-time management of traffic on arterial roads is to better manage traffic signals. In most cities, these are the only actuators available (some regions also have differential pricing schemes). Today this is done using several techniques starting with the timing of the individual signal, which may be based on a fixed timing plan or one fed by actuators (usually loop detectors) upstream of the signal. The timing plan can be updated to deal with different traffic flows at different times of day or based on real-time traffic data for the area. In some systems, signals are coordinated to support green waves (i.e., traveling at a set speed will result in a succession of green lights) possibly from a central facility or Traffic Operations Center (TOC). In most systems today these operations are based on a very small set of sensors, generally loop detectors and possibly cameras associated only with the most major streets.

The system is data-poor. Probes can provide the data to reduce delay and improve the throughput of individual signals. Of particular value is the information on where cars are headed and when they will arrive at a signal they are approaching. With this sort of information the signal can be ready for the majority of cars and let them pass through without slowing. Working in cooperation with intelligent vehicles, cars might be grouped so as to efficiently service one group, and then change the signal phase to service a second group just as it reaches the light. The vehicle group would know it would be serviced, reducing the need for long transitions between signal phases.

An even simpler use for probe data is to better manage the queues that are waiting for service from a signal. Currently there are only a few loop detectors at even the best-equipped

intersections, and these are not sufficient to provide detailed queue length measurements. With enough probe vehicles in the queues (especially when combined with arrival time information and some information on vehicle density) throughput at the signal can be increased. Estimates of the penetration required for active queue management are 30% and higher (Cheng et al. 2011). This is a serious drawback, especially if GPS probes are used. Some of these applications may be possible with cellular probes, which may have penetrations on this order, especially if one envisions an application that, when running, would help to provide better service at the light, thus, on average, decreasing delays for the user.

At much lower penetration rates ($\sim 10\%$) it is possible to determine the general traffic situation on arterial roads and use this information to invoke the appropriate timing plan for the system.

2.1.4 Weather

As mentioned previously, vehicles (and some phones) have a wealth of sensors that can be collected as probe data. Weather is one of the targets for these augmented probe systems. Many probes, phones and cars, have temperature sensors, and cars may have data on air pressure, precipitation, road surface, and friction coefficient with the road. Weather has a huge effect on surface transportation, and is of interest to drivers, but also has significant value in advanced vehicle applications in determining visibility and the driver's ability to see obstacles, estimating stopping distance, and the potential weather degradation to onboard sensors (cameras, LIDAR, acoustic). The best weather data today is largely useless for these purposes since the data is spatially very coarse, and most weather data is from well above ground level (sensed from distant weather radars). This data cannot determine if the road surface is wet (if the rain reaches the ground, or if the surface has dried), and may miss a local layer of ground fog (<http://www.weathertelematics.com/>).

The weather data from any individual vehicle is poor, but in aggregate, the data can be extremely robust. Access to the data is the critical issue, as the weather data is not generally available to the navigations system or other systems with access to wireless communications. Some weather data is standard on the OBD bus, other data may be available, although the format for this data may vary and be proprietary to the vehicle manufacturer. Several fleet management companies have developed connectors for the OBD port. These connections provide emissions data and support eco-driving applications, but, in research applications, are being used for weather data collection (http://www.its.dot.gov/connected_vehicle/road_weather.htm).

2.2 Historic Data

Real-time data can provide an intelligent vehicle with information on the immediate conditions, but it is not very useful, on its own, in predicting the future, even for very

short time periods. Predictions of a vehicle's future state (speed, location) can be done to some extent, based on infrastructure data (road geometry, speed limit), but these human behaviors are very hard to predict, especially when the vehicle systems do not have all of the information available to a driver. Humans respond to qualities of the road surface, sightlines, pavement markings, etc. in ways that are currently not quantifiable, even if the data was available in a digital map or to a vehicle sensor, which it is not. Advocates of autonomous driving often say that these tasks can be performed better by computers. This may be true, but, at least for now, is irrelevant since the autonomous systems must work on roads with human drivers, and have some ability to predict their actions. Drivers and passengers also expect a certain "human" feel from driving and are unlikely to accept a vehicle that handles differently. In order for Google to make their autonomous cars feel normal, they had the computers watch human drivers and learn to replicate their driving techniques (Vanderbuilt 2008).

Historic patterns from potentially tens of thousands of vehicles at a specific spot provide accurate statistical measurements of vehicle behavior. Measurements include averages, as well as the distribution across all relevant populations. Thus, approaching a curve the historic probe data provides the average speed of cars entering the curve, along with the speeds of the, say, 20th and 95th percentiles; approaching a stop sign the distribution of locations vehicles begin to slow is available. With sufficient data, behaviors can be sorted by time of day, weather conditions, traffic patterns, vehicle type, or any other variable associated with a probe vehicle. Particular drivers can be characterized against the probe-derived distributions, including which section of the distributions they occupy. This is another approach to predicting an individual driver's behavior, defining the envelope where they tend to drive, and determining personalized thresholds for information and warnings.

The effectiveness of these applications is determined by the total volume of data available, as long as conditions have not changed appreciably. Low penetration rates can be mitigated by longer collection times.

2.2.1 Transportation Network Design and Maintenance

Probe data is used to provide historical information on exactly how the road network works, how it fails, where it fails, and, to some extent, why it fails. Patterns of congestion are apparent, and with enough data these can be monitored throughout the day, correlated between days, with the weather, with incidents, and with events.

Historic data is one of the major requirements for designing a network – specifically information on where people travel and when. This data is key for feeding transportation models predicting the effects of a new road, signals, or other changes to the infrastructure. The probe data can also be used after the fact to validate the models, and prove that the work did what the advocates claimed.

Patterns of speed changes, hard braking events, and swerving can also be used to identify "blackspots," that is locations with a higher than normal incidence of crashes.

This data can, in turn, be used for routing vehicles around these difficult areas, or for tuning onboard sensors and algorithms.

Probe sensors can also be used to help characterize and maintain the road network. In states with snow, vehicles can be used to detect where the roads are still icy or need to be cleared based on wheel skid sensors. This can help to direct snowplows and make the most efficient use of sand or salt.

Boston has deployed a novel probe-based approach to finding potholes in the roads (<http://www.newurbanmechanics.org/bump/>). They have provided an iPhone app that detects the vertical acceleration of the phone in a vehicle as it crosses a pothole. This data is then reported to the city for analysis eventually leading to repair. In addition to the cost benefits of not having to spend time searching for potholes, this provides a generally unbiased metric for the overall state of the roads, which can be used to justify funding and to address political concerns that one area's potholes do not get as much attention as another.

The location of potholes, as well as speed bumps and rough roads, are used by intelligent vehicles to tune the suspension appropriately for the road ahead. Since it takes several seconds to modify the suspension tuning, this information must be provided from a map-type application rather than sensed in real time.

2.2.2 Traffic Control Attribute Detection

On many urban trips, the effects of signals (stop signs and stop lights) dominate the variance in travel time and total acceleration and deceleration. Knowing how traffic controls affect any particular route is critical to minimizing time, energy, or emissions (generally much more so than gradient). Historic probe data provides a way to identify the traffic control at a particular intersection and to characterize the likely cost in time, emissions, or fuel.

► [Figures 47.1](#) and ► [47.4](#) both show signatures characteristic of traffic signals, notably the stopping pattern before the intersection and the distribution of delays. Another way to look at the data is based on speed versus distance plots as shown in ► [Fig. 47.5](#). The presence of stop signs and stop lights and their impact on travel times can be determined by this method (Pribe 1999).

Metering lights for freeways are another significant influence on some trips. Probe data can be used to determine the usual operating times of the lights, as well as predicted delays when operational (real-time data can be used to determine current delay), as shown in ► [Fig. 47.6](#).

In many cities the main traffic corridors are not linked to a real-time management system, but often, these corridors will be timed for a certain speed or “green wave.” The lights are synchronized (possibly to a clock) so that vehicles traveling at a certain speed (and at a certain phase in the signal cycle) will hit a sequence of green lights allowing them to quickly travel the corridor. The synchronization of these lights is critical. The accuracy of timing coordination can easily be determined by using probe data. This can be used to

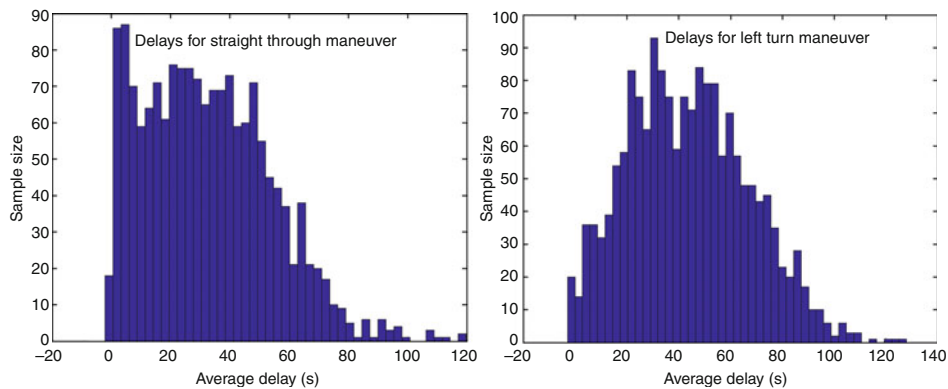


Fig. 47.4
Shows the delay profile for two different maneuvers at a very complex signalized intersection. The *left figure* is a straight through maneuver showing the bimodal distribution based on first encountering a green or red light. The *right hand graph* shows delays for the left turn across path maneuver (Courtesy of TomTom, © 2011)

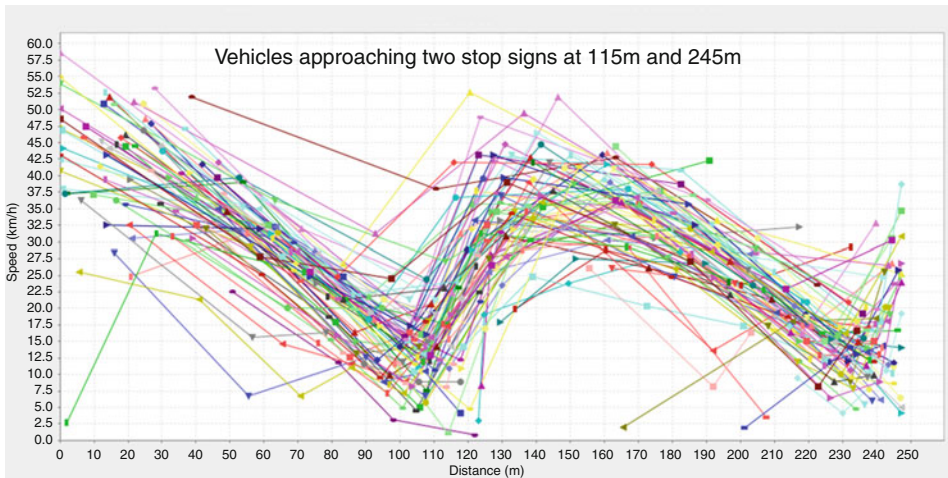
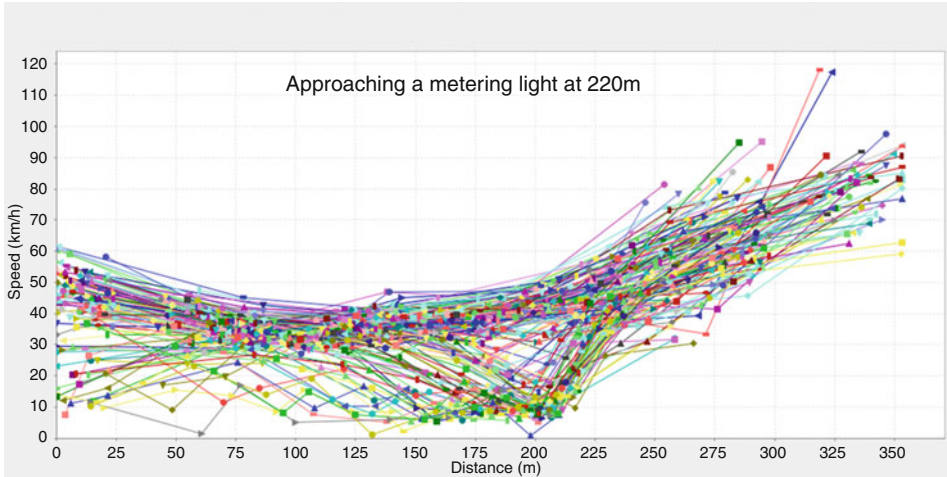


Fig. 47.5
Illustrates speed data for vehicle approaching two stop signs. This data can be used to determine the presence of the traffic control, as well as the expected location for a vehicle to start decelerating, and the distribution of initial deceleration locations. These facts are useful for stop sign warning systems. Data collected at 5-second increments and clipped at 10 km/h (Data courtesy of TomTom, © 2008)



■ Fig. 47.6

Shows several probe traces at a metering light on the on-ramp to a freeway. Some vehicles pass straight through when the light is off, others have a wait that can be observed in other views of this data. Data is at 5-second increments and truncated below 10 km/h (Data courtesy of TomTom, © 2008)

determine when the lights need to be re-synchronized, and if the work has been successful. This data is also valuable to a vehicle driving the corridor and attempting to catch the green wave.

This approach can be used to monitor any infrastructure project. The probe data “before” can be compared to the probe data afterward to determine quantitative improvements. In particular, this information can be used to monitor the delays of intersections and the travel times along corridors.

2.3 Probe Maps

There have been many projects to build road maps from probe data (Wilson 1998; Cao 2009) with varying levels of success. The purchase of Tele Atlas by TomTom was justified partially based on this approach, and has shown results in TomTom providing millions of map edits based on probe sources. But there are also difficulties with this approach since probe data maps drivers’ behavior, not the physical world like traditional maps. Physical characteristics can be inferred from behavior, but, it is just as hard to infer the physical reality from behaviors as vice versa, as discussed previously. For many applications, differences between the behavioral maps and physical maps are negligible (such as navigation) but the differences have significant import when used in some intelligent vehicle applications.

Building a probe-based map relies on the fact that data for an individual vehicle reflects the path of that vehicle to within the positioning error. Combining the paths of many vehicles on the same road can determine the location of the road to arbitrary accuracy, scaling roughly as the square root of the number of traces. This approach is being taken by several companies, most notably TomTom and Waze, to refine and build maps of the road network. These maps have several advantages over traditional cartographic methods (mapping vans, photogrammetric, survey), assuming a sufficient quantity of probe data (● [Table 47.4](#)).

The latency of probe based maps can be on the order of the latency of the probe data depending on how the trade-off between accuracy and reliability is made. In addition, since the map is built based on statistics, it is not prone to human errors in the map building process. For specific applications and business models there is an optimum map building solution that may combine both approaches.

Many of the attributes found in traditional maps can be derived from probe data.

- Road presence (limited lack of presence)
- Road geometry (including changes)

■ **Table 47.4**
Characteristics of probe-based maps

Parameter	Pros	Cons
Latency	On the order of a few to many times the revisit rate to a particular road. Can be on the order of minutes	Very little control of the sampling of particular roads. No way to prioritize certain areas
Accuracy	With enough data, can be arbitrarily high, as long as there are no systematic biases in the data	Accuracy takes time to achieve, and is largely based on the number of probes at the location. Cannot be easily controlled
Error sources	Process can be automated so that human errors are largely eliminated. All zero mean errors are averaged out	Grouping probes into like populations can be difficult and population errors may lead to application errors. Positioning errors are difficult to characterize
Accuracy attribution	Provides accuracy data for each point based on measurement distributions	Adds significant size to map database
Attributes	Expected driver behaviors are easy to capture, along with the distribution of behaviors. Roughly, but not exactly corresponding to physical observations	Does not directly provide a traditional physical map of the positions of physical things, such as paint on the road and road furniture. Requires inferences for physical features
Costs	Can be quite low, depending largely on communications and processing costs, assuming sources can be identified	Collecting excessive data over (expensive) wireless links can ruin business model

- Direction of travel
- Turn restrictions (including time of day)
- Gradient
- Traffic controls (as discussed above)

However, the probe-based behavioral map does not represent the same data as the physical map. The centerline of a lane in the probe map represents where cars drive. This is generally not the point halfway between the lane markings (which is the specification for physical maps), especially on a road with a cliff to one side or sharp turns. These offsets can be corrected to some extent, but for an autonomous driving system trying to follow the centerline, which represents the centerline choice of thousands of drivers, the behavioral map may be superior to the physical map.

Traditional mapping can collect house numbers, street names and signage, which cannot be collected by passive probes (at least without a camera onboard). This data is critical for navigation purposes, but probably not so valuable for many intelligent vehicle applications. Alternatively, probes collect speeds, stopping locations, delays and where someone actually finds a parking space, data not available from physical inspection of roads, and very valuable for some intelligent vehicle applications. The physical map can describe the curvature and bank angle of a curve, and a recommended speed can be derived from this data, but more valuable is the speed that thousands of drivers have chosen for this curve, especially drivers that handle other curves in a similar way. In addition, curvature and bank angle are difficult to measure and small errors can have relatively large effects on the recommended speed, whereas speed is native data for probe collections, and a statistical sample also comes with error estimates for each point in the database. Some inkling of the complexity of speed determination can be derived from an examination of the literature on traffic calming based solely on different paint patterns applied to a road surface (Ewing 2009). These patterns play to human perceptions of speed and distance, but are unlikely to have any effect on computer vision systems.

For some intelligent vehicle applications, probe-based maps provide entirely new ways to implement those applications. As an example, a parking space finder is difficult to implement if every parking space must be instrumented (or even mapped). In a behavioral system, this is implemented as a recommendation to follow a path historically used by the most successful parkers in similar conditions. The data required to support the behavior applications is generally easier to collect (with an adequate probe fleet) and will improve in a virtuous cycle as more probe data becomes available.

There are other applications where the relative merits of a physical map or a probe map are not so clear. These tend to be in instances where there are legal requirements associated with signage since signage is difficult to infer from probe data. Speed limits are a good example. At most locations the legal speed limit is very clear, yet very few people drive at the limit. For a fuel economy application, it is perhaps better to know the speed people actually drive to allow enhanced energy management on hills, yet it is still illegal to drive over the speed limit. Stop signs are another example. For most signs, most people stop and

the location can be derived from behavior. There are some stop signs where very few people stop, and they may not be detectable in probe data. Physical observation of a sign like this, likely on a minor road, will take significant expense, and these signs change rapidly. Warning for the sign might even be considered a nuisance by a driver familiar with the area. But running a stop is illegal, and occasionally, may have very serious consequences. Depending on implementation details of a stop sign warning system, completeness may be required or a hindrance.

Illegal behaviors need to be taken into account whenever regulations are inferred from behavioral data. In most areas and most restrictions, this is not a serious problem and the data can be calibrated to account for a minor population of illegal behaviors (e.g., wrong way drivers). This is a significant problem in some areas where regulations are generally not followed, but perhaps the accuracy of the data is not as critical in these locations where it is likely to be ignored.

2.3.1 Road Classification

There are also situations where the signage may be of less value than probe data, in particular road classification. Traditional road classification relies heavily on ownership, or a quick assessment of an expert passing by in a mapping van. Probe data allows for an assessment of how the road is used by people in the area: short or long trips, access to the freeway or the shopping center, only used during off peak hours, etc. This data is probably of more value to drivers than the type of shield on the road sign.


2.3.2 Intersection Geometry


In addition to the timing of the flows through an intersection, probe data can be used to reconstruct accurate geometry of an intersection. Several recent intersection collision avoidance projects in the USA and Europe (e.g., CICAS and SafeSpot) rely on precise knowledge of intersection geometry and the various vehicle paths. The concept of operations has this information coded into a map and broadcast to vehicles as they approach the intersection. This map is then combined with information on the signal phase and timing and used to determine appropriate vehicle movements and possible warnings. Critical to this concept is that an accurate map be provided, and one of the major failure modes is that the map is not synchronized with changes in the physical infrastructure. Building and maintaining the maps is a considerable expense for these systems. This can be done automatically from probe information. Since wireless communications is assumed, probe data can be sent to the intersection by departing vehicles. From this data, paths can be derived, and a map generated for distribution to vehicles. Once behaviors are associated with the various paths, these can be automatically mapped to the various signal states in the intersection. All of these steps can be performed automatically before the safety application is turned on. If, at a future time, deviations

from the historical pattern are found, possibly due to a change in the infrastructure, this process can be restarted.

This process relies on fairly accurate geometry, which may butt up against some practical limitations of current systems, although this could be remedied in any future intelligent vehicle deployment. The GPS systems in all probes to date are not intended for probe-based mapping. Most are intended for navigation and display types of applications where smooth behavior is a much higher priority than absolute accuracy, thus a filter is introduced to smooth GPS noise. This introduces a lag into the position data which will cause systematic errors in the position data from all vehicles turning through the intersection. These trends violate one of the fundamental premises of probe data mapping, that the observation errors are zero mean. The resulting map biases can be critical for intelligent vehicle applications where precise geometry is required, although they are negligible for navigation and traffic applications.

2.3.3 Instantaneous Speed Data

Current traffic products provide speed or travel times at a link level. This is adequate for routing, but not for many intelligent vehicle applications, such as the curve warning example provided earlier. Probe data can be used to derive the average instantaneous speed for all vehicles at a given spot, along with the distribution of speeds. An example of speed data leading into an exit ramp on a freeway is shown in  Fig. 47.7. The sample population can also be segmented by vehicle type, weather, time of day, or any other observable parameter. Furthermore, these speeds can be provided as a function of the route, not simply the location. For example, the speed for cars entering an intersection to go straight is likely to be higher than for cars, at the same point, that are planning to turn left. This data is simple to derive from probe data, and can be a critical differentiator for many applications.

Data such as that shown in  Fig. 47.7 directly supports several intelligent vehicle applications which are traditionally difficult to implement:

- Curve Speed Warning (determining recommended speed)
- Energy management (understanding the change in speed required ahead)
- Range calculations (including the total start-stop cycles on the route)
- Improved route guidance (based on differentiating routes based on speed offsets as well as lateral offsets)
- Driver feedback and training

The last item warrants further discussion. The behavioral data provides a normalized set of driver behaviors at all points on a road. An individual's behavior can be compared to the normal behaviors and deviations noted. These deviations can be used to identify opportunities for driver improvement that are within socially accepted limits, particularly improvement in green driving or safety. It should also be possible to derive safety or green driving scores, although there is no agreement on how to measure this score. The key to

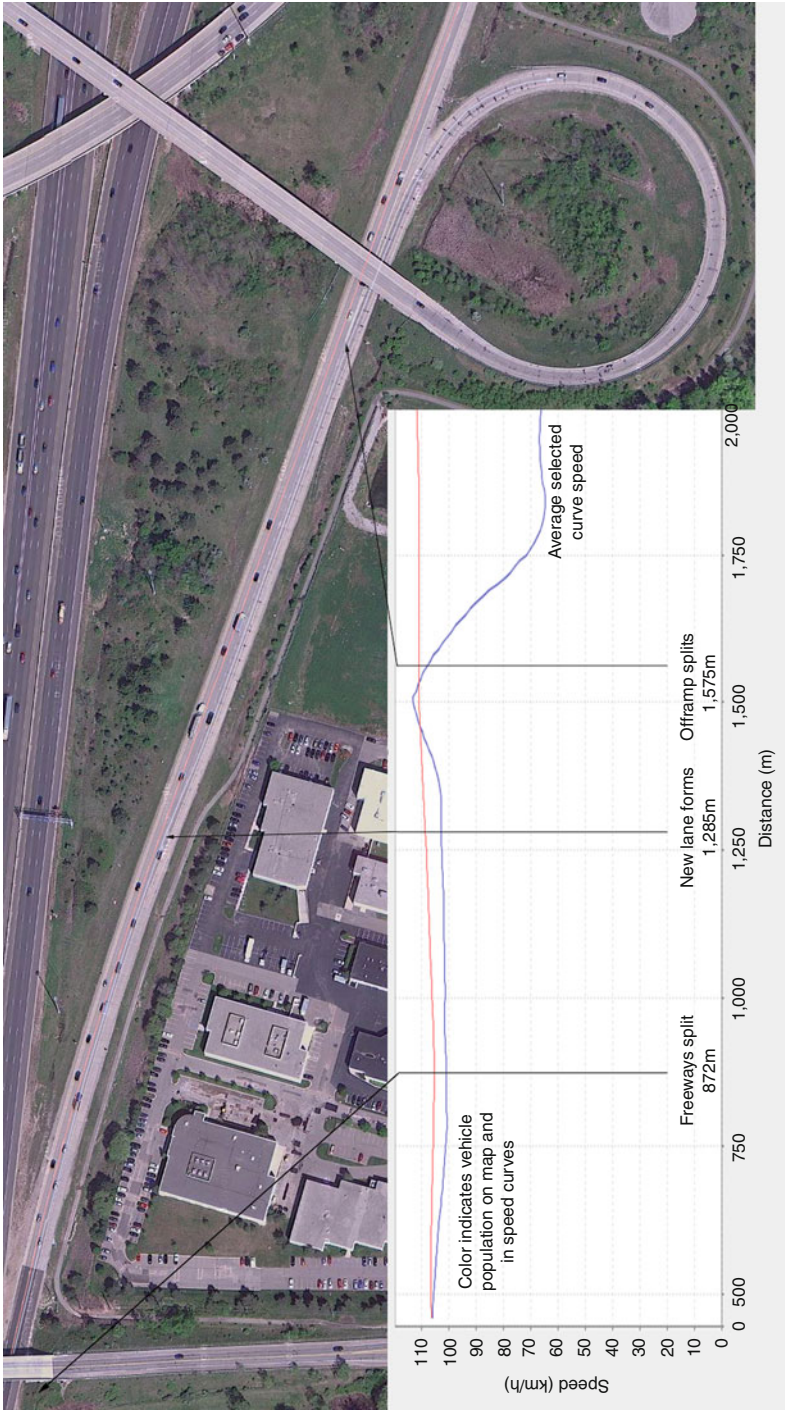


Fig. 47.7

Shows the instantaneous average speed map for two-vehicle populations on a freeway with the blue population taking the off-ramp, presumably after merging into the new lane as it forms. The red population stays on the freeway. An average speed for the off-ramp is easily determined from this data (Data courtesy of TomTom © 2008, imagery courtesy of Google)

effective feedback is to compare a driver with the population of other local drivers, not an imaginary driver from a Detroit simulator who precisely obeys the speed limit and approaches all stop signs in exactly the same manner.

2.3.4 Application Development and Validation

One of the most difficult aspects of developing advanced vehicle applications that take control of some aspects of driving is to replicate the feel of natural human behavior such as the proper anticipation of speed changes and the smoothing of curves. The feel is important for driver acceptance, yet, like face and speech recognition, which is easy for a human, this is difficult for a computer. This is due, in part, to the many subtle and subjective factors that a driver uses to determine speed, or exact lane position, including the tolerance for deviation from some idealized driving pattern.

For new intelligent vehicle applications, probe data is used to validate that an algorithm replicates the behavior of drivers as appropriate for many different roads. The first step in the evaluation of an algorithm is often to run simulations, using data from probes, and determine deltas between the algorithmic behavior and that of real drivers. The next generation of applications may be able to take behavioral maps as one of the inputs to the in-vehicle systems.

2.4 Privacy

No discussion of probe vehicles can ignore the issue of privacy. Probe data is generally used in the aggregate with relatively little value coming from individual vehicles which may deviate significantly from the normal behavior of interest, thus in its final form it is usually impossible to infer anything about a particular vehicle or driver. However, before the data is aggregated, the data exists on its own in the vehicle, probably through the communications network, and possibly on the servers that will eventually aggregate the data into an anonymous product. This is a problem for the most privacy sensitive. Encryption techniques can be used on the data in the vehicle, and throughout the process, but these increase the cost and complexity of the probe system, and may reduce the ability of a service provider to provide valuable services, including the primary service that the probe data supports.

One common approach to privacy is to delete the beginning and end of various trips (or better yet, not collect this data). This means the driveway where a vehicle spent the night is not obvious, but, depending on details, this may result in a significant loss of data. Most of this data will be on minor streets, which may not be important for traffic, but are critical for identifying parking locations and traffic controls such as stop signs, which are often found on these minor roads. Studies have suggested that in order to ensure privacy, a distance of over 2 km must be removed from the ends of trips (Krumm 2007). This is quite a significant amount of data, especially considering that many trips are not much longer than this.

Another approach used by many services is to not associate the series of location points from a vehicle trace (possibly by rounding the time stamp and ensuring a minimum volume of vehicles). This approach works reasonably well for traffic

applications, but is not acceptable for many applications where understanding the duration of stops and the details of various turning maneuvers is important. Many of the existing probe data sets, designed to support traffic, do not support these newer applications.

Like many situations around intelligent vehicles, the legal aspects of probe data privacy are unresolved. The question of who owns the data that might be stored in a car (as part of a probe collection effort, for accident reconstruction, or any other reason) is still undecided, let alone once the data leaves the vehicle and is given to a service provider. These issues are currently being addressed in the context of smartphone tracking, but are not likely to be resolved any time soon.

A related question is the suitable use of the data. Even the use of anonymous data has sensitivities. In the Netherlands, police used probe data to target areas where there was above average speeding for increased enforcement activities. The resulting public outcry caused the provider of the data, to issue an apology and promise that the data would no longer be licensed for such purposes. Of course, the same data could be used to raise the speed limit, to identify areas where traffic smoothing or other intervention is needed, or to target speeders in order to increase safety. This occurred at the same time as a public kerfuffle regarding location data collected by smartphones (early 2011), and legislation coming out of smartphone battles may inadvertently affect probe data collection.

Public perception of the use of probe data is likely to swing widely depending on the latest celebrity cases. In the short term, clear privacy and data use policies are the best defense. In the longer term, probe privacy questions are likely to be included in a larger discussion of right to privacy on the road. This discussion will be forced by mileage-based taxation, Pay-As-You-Drive insurance, and the movement toward autonomous vehicles that will require increased electronic visibility for vehicles and increased information on the roadways, which can only be practically obtained from probes. This is likely to be a long discussion, based on the lack of acceptance of relatively simple red light cameras (McGee 2003).

2.4.1 Increasing Probe Penetration

Probe data is here and it will not go away. Data from vehicles is beginning to become a commodity as more and more fleets have vehicle location systems and try to cut costs by reselling the data. Smartphones, tablets, and other mobile systems are also contributing data in support of very diverse applications. In some locations and times, a single aggregation system may have probe penetration of 15% of the vehicle fleet, possibly 5% in real time (penetration is difficult to determine and can only be done at specific locations) and there are many active aggregators.

High penetration of probes, along with new traffic management systems and algorithms, will significantly increase the capacity of today's physical road infrastructure. Work over the next few years will be on refining the existing models and learning more about the characteristics of probe data, learning to combine the disparate sources and increasing the penetration to enable new applications.

It is advantageous to combine the various sources of probe data into one large data set so that high penetration rates can be achieved. The framework for this is at the very early stages of discussion, but there may come a time when use of congested public roads requires that drivers provide data on how the roads are working, eventually leading to a true transportation network with infrastructure and vehicles determining the proper course of action in real time. In this scenario the anonymous probe data is available to a wide audience, and competition is based on the ability to derive information from the raw probe data, rather than simply access to the data.

3 Conclusion

Probe technology supports intelligent vehicles in two critical ways: (1) providing real-time information on the behavior of other vehicles in the region, enhancing foresight and situational awareness and (2) providing detailed, statistical information on historical behavior at every location, improving behavior prediction and personalization of driver interfaces. These are both important inputs to intelligent vehicle applications which are largely about intervening when driver behaviors are inappropriate for the situation. The increased knowledge about driver behaviors, and the expected range of deviations, can enable novel implementations of the applications.

Probe data collection systems are being widely deployed as the technology becomes more available, but the derived information products and applications are still in their infancy. Applications, other than traffic, are generally in the development stages. As the value of probe-enabled intelligent systems is proven, they will drive the collection of more probe data. In the near term, probe data will be collected primarily through cellular networks, but, DSRC systems are ideally suited to this purpose. The ability of DSRC to collect large volumes of vehicle information relatively inexpensively and often will help to drive their deployment. This is especially true for intelligent vehicles that need a lot of data (and have a lot of very interesting probe data to provide), and in locations where the infrastructure is actively managed. Probe collection will help drive the move to dedicated short-range communications systems.

References

- Cao L, Krumm J (2009) From GPS traces to a routable road map. In: ACM GIS'09, Seattle
- Cheng Y, Qin X et al (2011) Cycle by cycle queue length estimation for signalized intersections using sampled trajectory data. In: TRB 2011 annual meeting, Washington, DC
- ERTICO (2011) Successful conclusion of the CoCarX project. <http://www.ertico.com/successful-conclusion-of-the-cocarx-project/>. Accessed 25 Aug 2011
- Ewing R, Brown S (2009) US traffic calming manual. APA Planners Press, Chicago
- Krumm J (2007) Inference attacks on location tracks. In: Fifth international conference on pervasive computing, Toronto
- Kwon J, Petty K, Varaiya P (2007) Probe vehicle runs or loop detectors? Effect of detector spacing and sample size on accuracy of freeway congestion monitoring. Transp Res Rec 2012:57–63

- McGee H, Eccles K (2003) Impact of red light camera enforcement on crash experience. NCHRP Synthesis 310. Transportation Research Board, Washington, DC
- Meijer A, van Arem B (2012) Probe data from consumer GPS navigation devices for the analysis of controlled intersections. In: TRB 2012 annual meeting (Submitted)
- Pribe C, Rogers S (1999) Learning to associate observed driver behavior with traffic controls. *J Transp Res Board* 1679:95–100
- Vanderbuilt T (2008) Traffic: why we drive the way we do (and what it says about us). Knopf, New York
- Wilson C, Rogers S (1998) Potential of precision maps in intelligent vehicles. In: IEEE conference on intelligent vehicles, Stuttgart

48 Threat Model, Authentication, and Key Management

Stan Pietrowicz

Telcordia Technologies, Red Bank, NJ, USA

1	<i>Introduction</i>	1175
1.1	Vehicle Network Reference Architecture	1175
2	<i>Threat Model</i>	1177
2.1	Threat Agents	1177
2.2	Threat Motives	1178
2.3	Threats	1178
2.3.1	Compromise of Vehicle OBE	1179
2.3.2	RF Jamming	1180
2.3.3	RF Eavesdropping and Message Insertion	1181
2.3.4	Vehicle Tracking and Privacy Violations	1182
2.3.5	Access Point/Roadside Equipment Impersonation	1188
2.3.6	Viruses and Self-Replicating Worms	1189
2.3.7	Other Threats	1189
3	<i>Message Authentication</i>	1190
4	<i>Certificate Management</i>	1191
4.1	Combinatorial Certificate Management	1192
4.1.1	Privacy	1193
4.1.2	Effect of Certificate Revocation	1195
4.2	Short-Lived, Unlinked Certificates	1197
4.2.1	Privacy	1198
4.2.2	Authorizing Certificate Authority Functions	1199
4.2.3	Assigning Certificate Authority Functions	1201
4.2.4	Certificate Request Transaction Flow	1201
4.2.5	Detection and Removal of a Misbehaving Vehicle	1205
4.2.6	Weaknesses	1206

5 ***IntelliDrive Security Architecture*** 1207

6 ***Intrusion Detection in Vehicle Networks*** 1209

6.1 Intrusion Detection by OBEs 1209

6.2 Intrusion Detection by RSEs 1212

6.3 Network-Based Intrusion Detection 1213

6.4 Intrusion Detection by Certificate Authority 1214

6.5 Intrusion Detection by Application Server 1215


7 ***Conclusion*** 1215

Abstract: Security is an essential part of all vehicle networks. Communication among vehicles and roadside infrastructure needs to be secure, preserve vehicle privacy, and support efficient and effective removal of bad actors. The threat model for vehicle networks describes three categories of threat agents whose motives range from obtaining preferential treatment to tracking vehicles and disrupting transportation. Vehicle and roadside equipment, wireless communications, and network and software technologies are vulnerable to attack. The notion of privacy in vehicle networks encompasses the properties of anonymity and unlinkability. Vehicle tracking is a privacy threat that exploits vehicle communications, application transactions, and roadway conditions. Public Key Infrastructure is the predominant security architecture among vehicle networks, providing message authentication, integrity protection, and data encryption. The certificate management scheme affects privacy, the removal of bad actors, and system robustness. The combinatorial certificate scheme used in the US DOT proof-of-concept trial is an example of a shared certificate scheme. Removing bad actors in shared certificate schemes is challenging. Certificate revocation may affect many innocent vehicles, which may lose their network privileges. The short-lived, unlinked certificate scheme is an example of a unique certificate scheme that avoids the “one affects many” problem. It separates the certificate authority authorization and assignment functions and issues a large number of short-lived certificates, where certificate expiration may eliminate the need for revocation. Efficient and effective intrusion detection is critical to maintaining vehicle network integrity. Vehicle and roadside equipment, the certificate authority, application servers, and other network-based systems can participate in intrusion detection.

1 Introduction

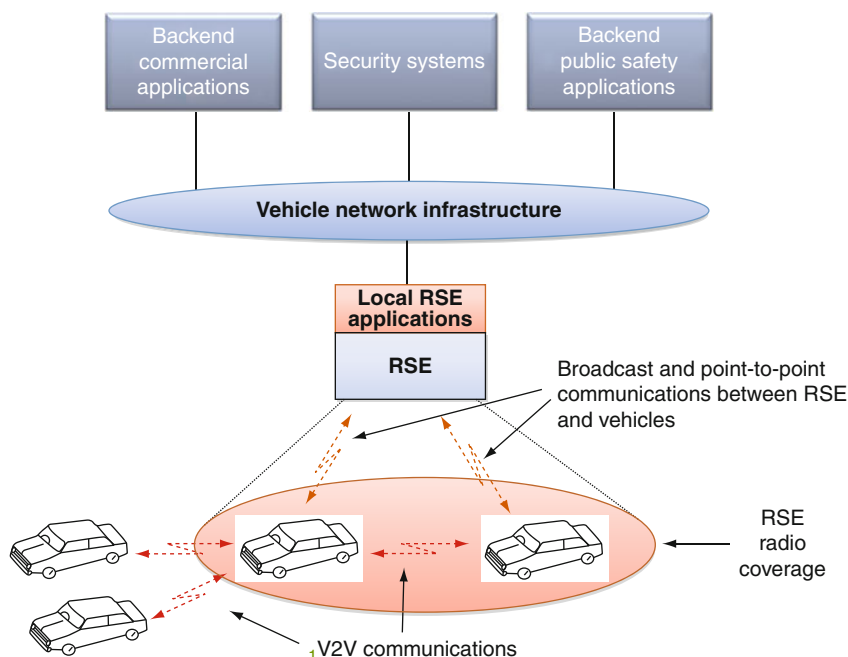
Security is an essential part of all vehicle networks. Communication among vehicles and with roadside infrastructure needs to be secure and preserve vehicle privacy. The successful operation of vehicle networks also requires means to efficiently and effectively remove bad actors. In this chapter, we begin by examining the threat environment for vehicle networks that support safety, mobility, and commercial applications. We identify three categories of threat agents and discuss several broad security threats to the integrity, confidentiality, availability, and privacy aspects of vehicle networks. We then discuss the security framework and mechanisms for message authentication in vehicle networks and contrast how two representative certificate management methods attempt to balance the complex problem of privacy and bad-actor removal. We present a high-level summary of the US DOT IntelliDrive security architecture. Finally, we close with a brief look at the topic of intrusion detection in vehicle networks.

1.1 Vehicle Network Reference Architecture

Without prejudice to any particular communications technology, a vehicle network will typically contain three types of communication endpoints as shown in  Fig. 48.1.

In order of greatest number, they are (1) vehicles, (2) infrastructure access points, which are also known as roadside equipment (RSE), and (3) back-end infrastructure servers. Although this may change as vehicle networks evolve, vehicles are generally network clients that run a multitude of safety and commercial applications and use network communication services. RSEs provide vehicles with the ability to access a fixed network infrastructure, where various servers for safety and commercial applications and infrastructure support services, such as security, reside. The back-end infrastructure will interconnect with the Internet to support in-vehicle applications, such as web browsing, and permit back-end data users, such as transportation agencies, to gain access to traffic data. RSEs may also support local services and applications, such as advertising network services and performing some latency sensitive services, such as toll processing.

Among these endpoints, both broadcast and point-to-point communications are used. Broadcast communications is by both vehicles and RSEs. Vehicles use broadcast communications for safety applications such as collision avoidance, where each vehicle periodically announces itself by means of a safety message to all nearby vehicles. RSEs also use broadcast style communications to advertise the network services available at an access point to all vehicles within its zone of coverage. Mobility applications, such as route guidance, and tolling use point-to-point communications between vehicles and RSEs or vehicles and back-end servers. Point-to-point communications between vehicles may evolve in the future vehicle networks.



■ Fig. 48.1
Basic vehicle network

The wireless technologies used in vehicle networks is still a topic of industry research. Present-day vehicle telematics services use 2G/3G commercial wireless networks and public/private WiFi networks to enable vehicles to infrastructure communications. Vehicle-to-vehicle (V2V) communications is latency sensitive, particularly for collision avoidance safety applications. At the time of this writing, the leading technology candidate is Dedicated Short-Range Communications (DSRC) documented in IEEE 1609.x standards. The system architecture of the IntelliDrive vehicle network sponsored by the US Department of Transportation (US DOT) is based on DSRC wireless vehicle communications with fixed and wireless commercial backhaul services from the RSEs. It is likely that future vehicle network architectures will involve a variety of communication technologies and a mix of dedicated and commercial wireless services, such as WiMax and the Long Term Evolution (LTE).

2 Threat Model

In this section, we identify and categorize the potential threat agents to vehicle networks and then examine the threats they pose in terms of motive, means, and opportunity.

2.1 Threat Agents

Threat agents are individuals or organized groups with the intent of abusing or performing malicious acts on any part of the vehicle network. Threat agents for vehicle networks are organized into three groups:

- *Category I:* Category I threat agents are solitary attackers who mainly operate on their own. They have fairly limited monetary resources and use the Internet as their main source of information. Examples of threat agents in this category are:
 - Unscrupulous or opportunistic individuals
 - Computer hackers
 - Automotive, electronic, or computer hobbyists
- *Category II:* Category II threat agents are corrupt insiders and small groups of individuals who are moderately coordinated, communicate on a regular basis, possess moderate resources, and can obtain information not publicly known or available. Examples of threat agents in this category are:
 - Corrupt employees of a vehicle network operator, automotive manufacturer, and automotive supplier
 - Internet-based attack groups
 - Unscrupulous businesses
- *Category III:* Category III threat agents are highly organized, focused, have access to expansive resources, can infiltrate organizations and obtain closely held secrets, and

may consider life and individuals expendable to achieve their goals. Examples of attackers in this category are:

- Organized crime
- Terrorist organizations
- Nation-state adversaries

2.2 Threat Motives

Threat agents are driven by a variety of motives whose severity and potential impact increases with the threat agent category. The basic motive for Category I threat agents is small personal gain or fame. Some of their interests in attacking a vehicle network are:

- Theft of service
- Avoidance of road tolls, congestion pricing, road pricing, parking fees, etc.
- Enhance personal privacy
- Commit insurance fraud
- Obtain preferential treatment from vehicle network
- Establish or enhance a reputation as a hacker
- Obtain sadistic pleasure in harming vehicles or disrupting a vehicle network

Category II threat agents are motivated by larger gains. Some of their interests in attacking a vehicle network are:

- Evade law enforcement
- Manipulate traffic authority decisions, for example, about traffic management or infrastructure improvements
- Redirect traffic to benefit a threat agent client
- Obtain significant personal gain
- Conduct corporate espionage or extortion

Category III threat agents are motivated by criminal, national, political, and terror agendas. Some of their interests in attacking a vehicle network are:

- Create an attack on a national infrastructure
- Perform an act of terror
- Conduct covert or intelligence operations
- Create civil, political, or economic disruption
- Use the system to benefit a criminal organization

2.3 Threats

This section describes several broad categories of threats to vehicle networks and discusses a few specific examples of each.

2.3.1 Compromise of Vehicle OBE

Since vehicles are the most numerous component in vehicle networks and the OBE installed in each is difficult to physically protect against attack, the compromise of OBE is a significant threat in vehicle networks. There are a number of ways to attack and potentially compromise OBE. One of the simplest forms of attack is to disable the OBE from participating in the vehicle network and cooperative safety applications. This can be done by grounding the OBE antenna, removing power from the OBE, or removing the OBE from the vehicle. Threat agents interested in this form of attack might be motivated by concerns over loss of privacy or that their vehicle is being tracked.

The next level of OBE attack is to swap or install OBE that belong to other vehicles. The purpose is to impersonate another vehicle or obtain a fresh OBE with valid credentials that can be used in an attack. OBE may be stolen from vehicles or obtained from salvage yards. Since the vehicle salvage process is not tightly controlled today, it is very difficult to disable OBE (potentially by remote means) or revoke their credentials immediately after a vehicle goes to salvage. Consequently, OBE sitting in salvage yards will be considered authenticated users of a vehicle network for a period of time after they have gone to salvage. In the worst case, an OBE in a salvage yard may maintain valid credentials up to their expiration date. To prevent such attacks, OBE and vehicles should mutually authenticate each other. Authentication failure on either side could be reported to back-end security monitoring systems as a security event.

A related OBE attack is to alter the sensor data and other information that the OBE uses in its application decision logic. For safety applications, this includes data from many of the vehicle's sensors, such as speed, braking, steering, and engine and body monitors, as well as other sensors available on the in-vehicle networks. For both safety and commercial applications, Global Position System (GPS) signals and inputs to the OBE navigation system for "dead reckoning" might be targeted.

Aside from the attacks on the OBE inputs and swapping of OBE units, the OBE hardware itself is subject to attack. The OBE is an embedded computer system with a processor, nonvolatile storage, such as FLASH memory, application firmware, configuration data memory, and a variety of input/output interfaces. The credentials and keys, particularly the OBE private keys in a Public Key Infrastructure (PKI) security system, need to be stored and protected in OBE hardware. OBE hardware attacks may attempt to extract the unit's credentials and identity information, alter the OBE configuration information, change the code logic in the application firmware/software, and add rogue software. If security controls such as firmware signing are used to protect application code, the secure bootloader that evaluates the integrity of the firmware image and OBE may be the subject of attack.

Debugging interfaces, particularly the on-chip JTAG interface, are a primary security concern in all embedded hardware systems. If not properly disabled or protected, threat agents can connect emulators to monitor program execution and processor state, alter OBE memory, retrieve firmware from FLASH for disassembly and analysis, overwrite FLASH memory to change code logic and credentials, and test potential exploits.

A successful compromise of OBE hardware will provide a threat agent with control of an OBE that is recognized as a legitimate and authenticated user of a vehicle network.

There are well-founded reasons to be concerned about the likelihood of a successful OBE compromise. Attacks on OBE hardware will occur with almost 100% certainty. Hardware security is a weak link in many industries today. Good hardware security practices are not often followed partly because they increase unit cost, there is a lack of security expertise in hardware design groups, and business decisions to avoid security controls that may prevent the manufacturer from recovering hardware for refurbishment or being able to recover from a technical mistake. The successful compromise of OBE credentials and firmware are two of the most significant threats to OBE.

2.3.2 RF Jamming

Vehicle networks inherently depend upon wireless communications to enable vehicles to communicate with other vehicles and fixed network infrastructure. Signal jamming is a threat in most wireless systems and, in this case, to the availability of services in a vehicle network. A threat agent can flood the spectrum used by a vehicle network with sufficient energy to degrade the carrier-to-noise/signal-to-noise ratio of vehicle and RSE transmissions to the point where receivers either cannot demodulate the signal or encounter numerous transmission errors that severely degrade network throughput.

Most vehicle networks operate in a narrow band of spectrum, which is either dedicated for their use or part of a commercial wireless communications service. Unlike WiFi networks, for instance, the spectrum used for vehicle networks is generally a licensed band. In the case of the US, 75 MHz of spectrum at 5.9 GHz has been reserved for vehicle networks. This is advantageous, if not required for vehicle safety applications, which require low latency and are by design intended to have priority over other vehicle communications. Narrow bands of spectrum are susceptible to flooding attacks that can overcome even jam-resistant technologies, such as frequency hopping and direct sequence spread spectrum. These technologies depend upon spreading their signal energy across a wide spectrum, which would require a large amount of RF energy to prevent signal communications.

Equipment to jam wireless signals can be readily acquired or built. In the case of DSRC, which is based on the popular 802.11 standard, some WiFi computer network cards could be altered or configured to transmit in the 5.9 GHz band. Indeed, this is how communications traffic monitors were built for use in some vehicle network trials. Threat agents can also compromise OBE and use their radio systems to jam the band.

Given the wide geographic area that vehicle networks cover, RF jamming would most likely be a localized event. A single malfunctioning vehicle or RSE transceiver or jamming device located at a fixed location or in a threat-agent vehicle would most likely disrupt communications within a coverage radius of a kilometer or less. Other parts of the network would continue to function. A threat agent might employ a jamming attack to disrupt communications in a targeted area by placing a jamming vehicle at the site.

However, a broad denial of service attack would require multiple jamming devices preferably with high-power transmitters to cover larger geographic areas. Such an attack could be implemented if OBE are compromised with a radio jamming worm/virus that spreads through the vehicle network. A complete network shutdown is unlikely, except possibly in a military attack.

2.3.3 RF Eavesdropping and Message Insertion

Unlike wired communications, where the transmission medium can be physically secured, wireless signals are difficult to contain and, consequently, access to the low-layer communications signal is generally available. Eavesdropping on wireless signals in a vehicle network is an achievable method of attack. It is made easier when the wireless technology used in a vehicle network is built upon a technology that is used for general-purpose wireless computer networks. This is the case for IntelliDrive, where a threat agent can construct an RF eavesdropping device using an ordinary computer, custom software, and flexible WiFi network card.

Unlike a traditional computer network, where eavesdropping on a link higher in the network structure may provide access to aggregated traffic, eavesdropping in vehicle networks will only provide access to signals in a single RSE zone, unless multiple tapping devices are deployed. Even if eavesdropping was performed on the backhaul link used by the RSE, it would at best yield access to some of the communications in the area served by the RSE, and unlikely to most of the broadcast and V2V communications that would not normally be routed to the back-end. Access to links that aggregate traffic in a vehicle network is more difficult than in an enterprise environment.

A common means to protect against eavesdropping on a link or network connection that cannot be trusted is to encrypt communications using a cryptographic cipher. Secure web transactions, for instance, are conducted using a secure session protocol called Transport Layer Security (TLS) that provides both confidentiality and integrity protection. However, not all vehicle communications will be encrypted in a vehicle network. While tolling, commercial applications, and security transactions in vehicle networks will likely use encryption, the computational cost and delay associated with encryption is an impediment for safety communications. Safety applications require low-latency communication at message rates of 10–50 times a second. Aside from delay, decrypting safety messages from a multitude of vehicles presents a resource issue. As a trade-off, safety messages, which are meant to be available to all authenticated entities in a vehicle network, are typically not encrypted. Without encryption, their payloads are also available to threat agents and unauthenticated entities who can eavesdrop on wireless signals. Consequently, the content of communications that are not encrypted needs to be carefully controlled to limit the security risk.

Once a threat agent has access to a network and can eavesdrop on communications, a natural next step is to insert phony messages and replay messages that have previously captured. Virtually, all vehicle communications require authentication in the form of

message authentication, secure broadcasts, or secure sessions. While safety messages do not have confidentiality protection, they are more importantly integrity protected and authenticated, typically using a digital signature. Means used in vehicle networks for authentication are discussed later in this chapter.

2.3.4 Vehicle Tracking and Privacy Violations

Two of the greatest concerns in vehicle networks are the ability to track a vehicle and commit violations of privacy. Each vehicle that participates in a vehicle network makes itself present, even if just anonymously, by the fact that it generates wireless communications that may be subject to eavesdropping. Some information about vehicles that participate in a vehicle network will be collected, and pseudonyms and possibly identifying information will be stored and used by systems and applications. The goal is to minimize or eliminate the use of identifying information and any information that can track a vehicle. Preventing vehicle tracking and preserving privacy in the face of security requirements to authenticate every message and efficiently remove malicious actors without impacting innocent vehicles is probably the most complicated technical challenge in vehicle networks.

A Notion of Privacy

In vehicle networks, two properties define privacy:

1. Anonymity
2. Unlinkability

Anonymity is the inability to identify or enable identification of a vehicle, its owner, or occupants as a result of its participation in the vehicle network. This includes, but is not limited to, DSRC message communications and information processed or retained in any part of the vehicle network. Identifying a vehicle means to obtain one or more distinguishable vehicle attributes that can be definitively linked to the vehicle owner or vehicle occupant.

Unlinkability is the inability to definitively associate observations, data, or information, such as vehicle communications, as belonging to a particular, but possibly unidentified vehicle, vehicle owner, or occupant. Unlinkability implies the inability to track a vehicle's path, especially as it moves from one RSE zone to another.

Anonymity and unlinkability are distinct but closely related components of privacy. Often, initial attacks on privacy focus on immediately compromising anonymity and identifying a subject of interest. However, if those efforts fail, other techniques, such as traffic analysis and correlating various sources of information, can be used in part of a multistep process that can slowly and patiently resolve the identity of a subject. For instance, without knowing its identity, an anonymous subject can be studied to understand its behavior and anticipate its moves. It can be characterized so that it becomes highly distinguishable in a crowd and can be tracked. Tracking is a very powerful technique. It allows a threat agent to keep a record of where a subject has been and follow it to where it is going. Each of the events that form part of the subject's path

provides an additional opportunity to identify it. Various sources of information that on their own might be considered anonymous could be correlated together to narrow or resolve an identity. If a subject can be tracked, its anonymity is at risk, and the additional step to identify it may only be a matter of time. For instance, an anonymous vehicle that is being tracked may divulge its identity when it passes in front of roadway camera that can record its license plate. Conversely, it is also possible to identify a subject at a particular point in time, but not be able to track it.

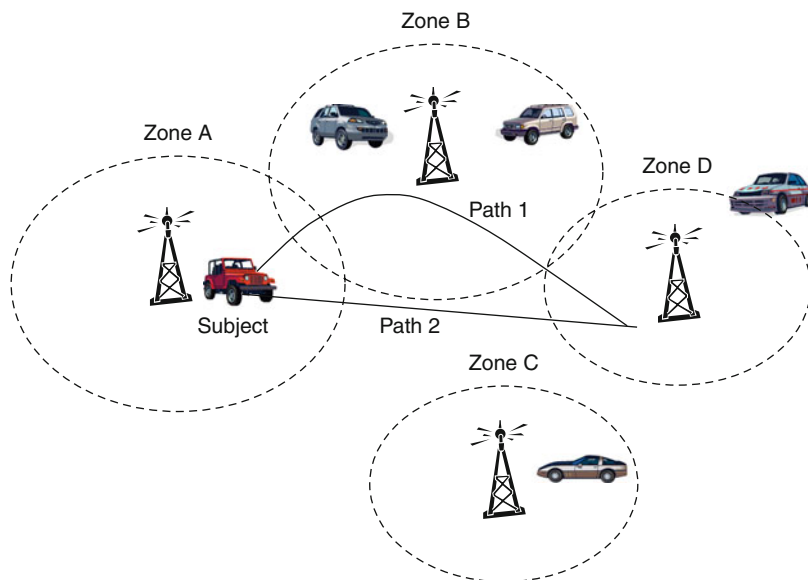
Anonymity and unlinkability are of greatest concern for safety applications. Unlike private applications for which a vehicle user has the option to subscribe or decline a service, mandatory safety applications are envisioned to be an intrinsic capability in all vehicles without consumer choice. As a result, there is much concern about maintaining vehicle privacy and not violating vehicle owner civil liberties, which could greatly affect public acceptance of vehicle networks.

Privacy is best preserved when a threat agent cannot identify a subject, cannot link its communications, and cannot track its movement or anticipate its actions. Nonetheless, there are practical limits to privacy. A person who owns a vehicle will drive, for the most part, on public roads where they can be observed, monitored, and possibly tracked without ever participating in a vehicle network. Thus, a reasonable approach is to preserve the level of privacy that exists in the absence of a vehicle network when participating in a vehicle network. For more discussion on privacy in vehicles networks (see Di Crescenzo et al. 2010).

Vehicle Tracking

The problem of vehicle tracking can be divided into two parts. The most significant concern is when a vehicle can be tracked from one RSE to another, thereby establishing its path of a significant geographic area. Without having a large number of eavesdropping probes, the best candidate to track vehicles is the vehicle network infrastructure itself. If applied differently, the RSEs can be monitoring points to support an application that can track a vehicle as it moves in the network. A second and lesser concern is tracking a vehicle within an RSE zone. Since RSE zones are not very large and session oriented communications need to maintain some identifiers in order to successfully engage in two-way communications, tracking within an RSE zone may be a practical limitation.

To track a subject vehicle, one does need to know its identity, but it is necessary to be able to distinguish it in a crowd by one or more discernable characteristics. ♦ *Figure 48.2* illustrates the task of tracking a subject vehicle. First, it must be determined that a subject vehicle is in a particular zone, for example, Zone A and some information about it usually needs to be collected. Second, vehicle transmissions in Zone A and the surrounding zones need to be monitored for an event that can possibly indicate that the subject vehicle has crossed zones. If Path 1 in ♦ *Fig. 48.2* is taken, the subject vehicle is always within radio range of the system. As it crosses between Zones A and B, communication is not interrupted and the transition of leaving one zone and entering another is relatively quick. Static identifiers and temporal relationships in messages observed by RSEs can be used to track vehicle movement across zones.



■ Fig. 48.2

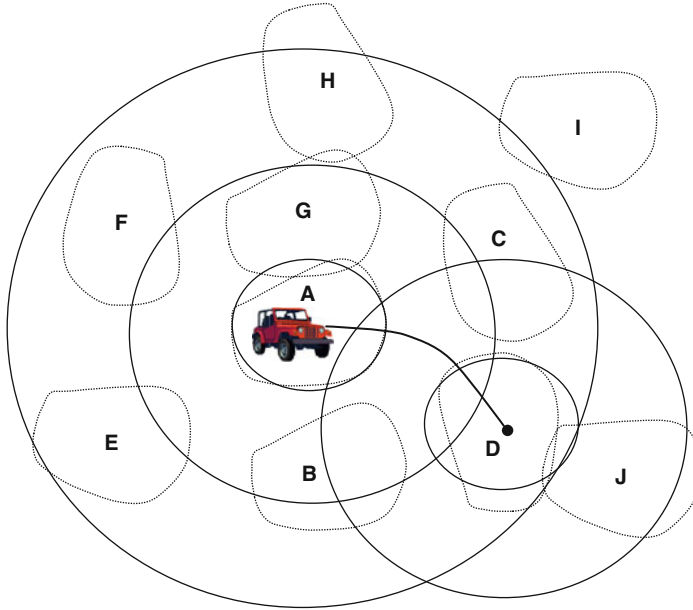
Tracking a vehicle across radio zones

If the vehicle takes Path 2, it traverses an area with no radio coverage. Using knowledge of the radio zone maps and vehicle position and speed, it might be possible to predict when a subject vehicle is about to leave radio coverage and where it might turn up. Temporal relationships with messaging in Zone A, for instance, become stale as the distance between the two zones becomes greater. Unlike Path 1, the transition in Path 2 out of one zone and into another is slow, which makes tracking more difficult. During the transition, other vehicles could enter Zone D and obscure the true point of entry for the subject vehicle.

Consequently, successfully tracking a subject vehicle in a system designed to protect its privacy has a number of uncertainties. Graphically, this uncertainty can be depicted as a series of concentric circles of confidence overlaid on an RSE radio zone map of the area as illustrated in ● Fig. 48.3. The innermost circle, for instance, over Zone A indicates the area of highest confidence where the subject vehicle is most likely to be located. When the vehicle moves to another zone, it enters an outer ring of lower confidence, until such time as one or more events sufficiently distinguish it among other vehicles. For instance, if the vehicle moves into a zone with low vehicular traffic, confidence in the zone where a vehicle may be located could suddenly increase. At some threshold of confidence, the concentric circles refocus on a new zone as shown in ● Fig. 48.3.

Below is a list of other factors that can increase the chance of successfully tracking a vehicle:

- *Overlay a Roadway Map:* Tracking a vehicle's movement across zones could be further improved by overlaying a roadway map on the RSE radio zone map. Most vehicles will



■ Fig. 48.3

Circles of confidence in tracking

remain on public roads, which occupy limited and known paths through radio zones. There are a finite number of ways in and out of zone. Distances, speed limits, and physical laws could be used to estimate time of transit. Knowledge of the area and even traffic alerts from the vehicle network can be used to help track a vehicle's movement.

- *Low Vehicular Traffic:* RSE radio zones that normally have low vehicular traffic or periods of the day when traffic is exceptionally light reduce the number of vehicles from which the subject vehicle needs to be distinguished. In the best case, the subject vehicle is the only vehicle in a radio zone, which makes it the only possible source of vehicle transmissions.
- *Positioning Data from Safety Messages:* Any positioning data that is available from vehicle transmissions, such as safety messages, will help determine its location. Position data provides location resolution within the zone for a vehicle source. Multiple messages with positional data from a vehicle source would establish the vehicle direction. One or more messages with position data could also indicate the next most likely zone that a vehicle will enter.
- *Correlation with Other Traffic Systems:* Information can be correlated from other traffic systems, such as traffic cameras, to better track, periodically observe a vehicle, and possibly make a positive vehicle ID.
- *Radio Frequency (RF) Signal Tracking:* The vehicle is, to some extent, a radio beacon. RF signal strength, unique characteristics of the vehicle transmitter, Doppler frequency shifts, and other such physical layer characteristics can be used to help track a vehicle.

Tracking Using Application-Specific Identifiers

There are at least three different groups of applications where significant vehicle tracking and privacy concerns exist. The first group involves safety applications. At least one or more safety applications will be running whenever a vehicle is operated. Cooperative safety applications require vehicles to periodically transmit “heartbeat messages” that contain information about the vehicle and its position, direction, and speed. Each vehicle generates heartbeat messages at a nominal rate of about 10 heartbeats/s, which makes every vehicle a broadcast beacon. Since the benefit of cooperative safety applications can only be achieved when there is a high penetration of participating vehicles, safety applications may become mandatory, exacerbating the concern of vehicle tracking. Because safety messages are not encrypted and use broadcast communications, their payload is visible in clear text and each message bears some form of authentication. Any unique or locally unique identifier or pseudonym that appears across messages creates linkability and presents an opportunity to track a vehicle. This applies not only to the safety message payload, which does yield position and time information, but it also applies to the security information and lower layers of the transport technology. For instance, a signed safety message in a PKI-based vehicle network will bear a certificate, certificate ID, and public key that need to be used to verify the signature. Similarly, an IP address that is assigned when a vehicle enters RSE coverage and is used by a vehicle for all its communications creates linkability. The Media Access Control (MAC) address, which is generally a globally unique number used at the link layer, will also establish linkability.

One method to reduce message linkability is to randomly choose a new MAC address and IP address when entering an RSE zone. Time, distance, and message count triggers for changing these values also need to be considered because RSE zones may not be geographically contiguous, but instead be separated by a significant distance. If time and distance triggers are not considered, a vehicle may maintain its pseudonyms well outside RSE zone coverage.

Private and commercial services are a second group of applications that create concern with respect to vehicle tracking and privacy. Unlike safety applications, the communications exchange for commercial applications is likely to be two-way and may need to extend across multiple RSE zones. These services are likely to be transactional in nature, i.e., have multiple requests and responses, and require reliable and secure transfer. Commercial applications are likely to employ application layer security to authenticate vehicles and services and create secure sessions. Techniques commonly used to reestablish rather than initiate a new secure session when entering an RSE zone might provide an eavesdropper with a way to associate session communications. In these communications, the potential exists for linkable identifiers or pseudonyms to be present, which can be used to track a vehicle's path.

Some of the identifiers or pseudonyms that may assist in tracking a vehicle include:

- Destination IP addresses of application servers or proxies.
- Message authentication mechanisms, such as X.509 certificates. Unlike safety applications where anonymous certificate management methods have been considered to protect anonymity, commercial applications, if they follow their Internet models, would assign unique credentials to each vehicle.

- Session IDs and credentials used in secure sessions.
- Service identifiers and pseudonyms in the application payload.
- Identifiers in the reliable transfer protocol.


Another concern is the potential cross-linkage between safety and commercial applications. Even if safety applications are completely unlinkable on their own, a commercial service that exhibits poor unlinkability in its communications may be used to reestablish linkability with safety messages, which may provide definitive vehicle position information. In this fashion, vehicle tracking alternates between monitoring safety messages while in an RSE zone and using the communications of commercial services to reestablish the trace and association with safety communications when the vehicle enters a new RSE zone.

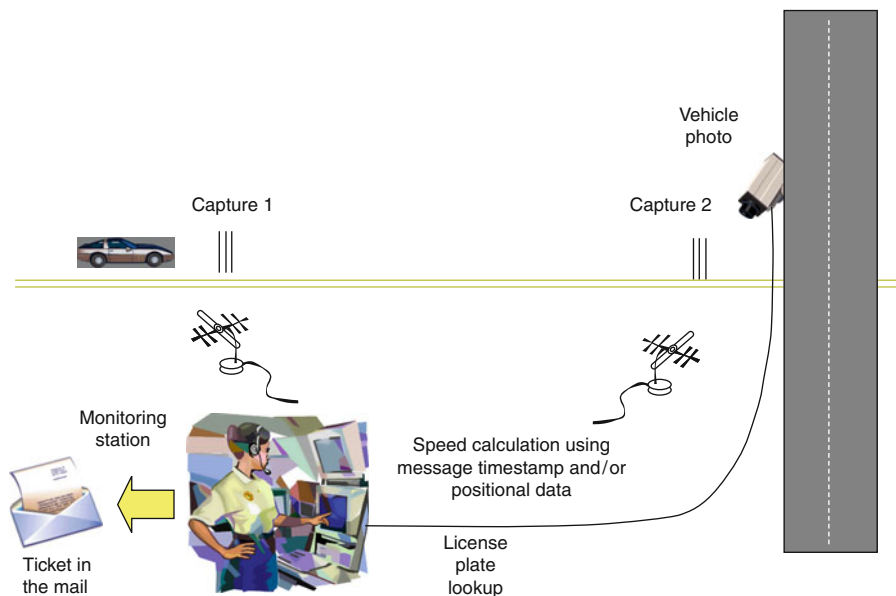
A third group of applications for which vehicle tracking and privacy concerns exist are those that collect information about a vehicle and its location. Two examples are road tolling and probe data collection. Whereas the previous set of applications might be more applicable to real-time tracking, these applications provide information that is better suited to establishing vehicle path in post-event analysis. Road tolling based on fees collected at specific points along a roadway establishes a vehicle at particular location at a particular time. Because tolling requires the billing of fees, positive vehicle identification needs to be made. Of course, with fixed point tolling, vehicles can only be tracked to a tolling point. However, new road-pricing technologies will be able to track a vehicle using GPS and potentially apply toll charges on every road traveled by a vehicle.

Probe data collection is a traffic data collection service whereby a vehicle application periodically collects information from the vehicle as it travels a road and records events, which are uploaded to a back-end application in the vehicle network. Probe data management applications can actually instruct the vehicle as to what information to collect. The information is meant to help traffic management centers better understand traffic patterns, road conditions, weather, problem areas, and influence route guidance decisions. While the intended use of the information collected is not to track a particular vehicle, tracking is a concern if the safeguards are not employed to anonymize and aggregate the information.

Considering all three groups of applications in terms of their potential impact on privacy, a major difference between commercial applications and mandatory safety or traffic data applications is the driver's right to opt in or out of a commercial service and the need to get approval for certain information to be collected and how it will be used.

Speed Trap Example

The following example demonstrates how safety messages could be used to create a speed trap without requiring the permission of the vehicle network operator. Consider a speed trap composed of two vehicle-communications monitors situated along a stretch of roadway as depicted in  Fig. 48.4. Each monitor captures the safety messages transmitted by vehicles as they pass. The monitors are positioned close enough to each other such that the MAC address of the vehicle does not change between the two capture points. No exits or turnoffs exist between the two capture points. At the second capture point, a roadway camera is installed to record the vehicle image.



■ Fig. 48.4
Speed trap example

Assume a vehicle now travels into the area and it is generating heartbeat safety messages that contain timestamp and position information. At the first capture point, the vehicle's heartbeat messages are recorded. Sometime later, the vehicle approaches capture point 2, where not only are the vehicle's heartbeat messages recorded, but a photo of the vehicle showing the driver and license plate is taken. An average speed calculation is then performed as a post capture process either by using the timestamp and position information in the heartbeat messages or based on the known locations of the capture points and a timestamp made by the recording equipment. The heartbeat messages are matched based on the MAC address contained in each message, which remains the same during the period of time the vehicle travels this section of roadway. If the vehicle is speeding, an automated system performs character recognition on the license plate, looks up the owner in vehicle registration databases, and mails a summons to vehicle owner along with the photo. While a traditional radar system could have been positioned at capture point 2, the benefit of this system is that it measures average speed instead of instantaneous speed.

2.3.5 Access Point/Roadside Equipment Impersonation

Vehicles will be driven in areas for which the location of trusted RSEs is not known. Not all geographic areas will be covered by an RSE. In fact, many local roadways, except at high incident intersections, will not have coverage. A threat agent could set up an imposter RSE in an area without coverage and begin advertising services to vehicles. The threat agent is

then in a position to conduct a number of man-in-the-middle attacks should a vehicle attempt to communicate with the back-end systems through the imposter RSE.

In a related threat, a threat agent could physically move an RSE from one location to another, which may result in the RSE broadcasting incorrect information for its current position. For instance, in a speed warning application, the RSE may broadcast incorrect roadway geometry, which may cause driver confusion and accidents, especially if the OBE responds to it and provides erroneous alerts and warnings to the driver.

2.3.6 Viruses and Self-Replicating Worms

In any computer network, there is always a threat of viruses and self-replicating worms. Vehicle networks will support a variety of safety, mobility, and commerce applications. Depending on their purpose, these applications will use a variety of software technologies, including many commonly associated with the web, such as Java. In addition, many OBE use computing hardware and operating systems more commonly used for general-purpose computing. Some vulnerability and exploit techniques associated with general-purpose computer networks and web applications might also impact vehicle networks.

Vehicle networks, unlike traditional managed networks, might also be more susceptible to the spread of malicious code because of the uncontrolled nature of ad hoc networking that occurs among vehicles. In a fixed hierarchical, managed network, malicious communications can be detected by a trusted authority and potentially filtered or blocked. However, in vehicle networks, it will frequently occur that two or more vehicles that have not previously encountered one another will communicate outside the coverage of an RSE. Since vehicles will likely outnumber RSEs by almost three orders of magnitude and at least as much when considered in terms of geographic area covered by RSEs versus total roadway area, much of the communication in vehicle networks will take place in an unmanaged environment.

Another concern with the spread of malicious code is over-the-air code updates and, more specifically, the floating of code updates from one vehicle to the next. Over-the-air code update provides a mechanism for propagation of malicious code. The concern is compounded when each vehicle has the potential to become a source of malicious code for the rest of the population through ad hoc networking. As is the case for most vulnerabilities, this threat could be combined with others to create attacks with greater potential harm. For instance, a self-propagating virus might be built that takes control of the OBE radio system to perform a distributed jamming attack that covers a large geographic area, overcoming the limitations of manually creating a few jamming devices.

2.3.7 Other Threats

Other threats to consider in the security architecture for vehicle networks include:

- Certificate authority compromise
- Compromise of back-end system access

- Insider threats
- Supply chain threats (since the vehicle network at some point may be considered a critical national infrastructure)
- Use of vehicle network systems to track vehicles
- GPS signal starvation or tampering

3 Message Authentication

Secure vehicle communications is critical to the integrity of cooperative safety systems. Authentication and integrity of vehicle messages, the preservation of vehicle privacy, and the ability to efficiently remove bad actors from vehicle networks are three important and tightly interrelated components of vehicle network security. V2V communications used in cooperative safety systems is inherently an opportunistic exchange among strangers. Vehicles may have no prior contact with each other and may never have contact with each other again, but must be able to trust the information each shares. Each vehicle must be able to verify the integrity and authenticate the messages it receives from other vehicles before it can use the information with confidence to render safety decisions.

The security architecture for vehicle networks, such as IntelliDrive, is based on Public Key Infrastructure (PKI). PKI provides the cryptographic mechanisms for secure message authentication and encryption and a framework for the life-cycle management of security credentials that include a public and private key pair and a certificate from a trusted entity.

A fundamental reason for the use of PKI in vehicle networks is for message authentication using digital signatures. A digital signature is essentially a one-way hash of a message that is encrypted with the private key of the sender. The asymmetric cryptographic properties of the public and private key pair allow one key to be made public so that any recipient of a signed message can verify its digital signature. Successful verification of a digital signature attests to the integrity of the message content and, in combination with a signed certificate from a trusted entity, authenticates the source of the message. The inclusion of various attributes, such as expiration date and geographic area of validity, in the signed certificate provide a convenient means to apply additional authorization controls. PKI also provides for asymmetric encryption of information and supports the establishment of secure sessions using well-known protocols, such as Transport Layer Security (TLS).

In a vehicle network, each vehicle is assigned one or more certificates according to a certificate management scheme. A vehicle sends a secure broadcast of a safety message by signing the message using its private key. Any vehicle receiving the broadcast can validate the signature to confirm message integrity and verify that it came from an authorized source without having previously exchanged security credentials with the source. A Certificate Revocation List (CRL), which is also administered by the trusted entity, is used by recipients to confirm that the certificate is still valid and has not been revoked. Other security checks using the attributes in the certificate provide for finer authorization


controls. In this manner, PKI provides the basic integrity and authentication mechanisms to support secure communications between vehicles that are strangers to each other. PKI certificates and keys are also used by RSEs to secure broadcasts of service advertisements and prevent RSE impersonation. Use of geographic information in the attributes of RSE certificates prevents the operation of an RSE that has been moved without authorization. The IEEE 1609.2 standard (IEEE 2006) for wireless access in vehicular environments defines the use of digital signatures to authenticate V2V communication.

The above description is a simplistic view of message authentication in vehicle networks. PKI operations are computationally intensive and pose a burden to embedded systems, such as OBE. A vehicle may receive numerous messages that need signature validation. In a highway situation with each vehicle generating safety messages at 10 messages/s, a vehicle can receive upward of 1,000–2,000 messages every second and each could require cryptographic operations. To reduce resource requirements, secure broadcast protocols, such as Time Efficient Stream Loss Tolerant Authentication (TESLA) are under investigation by industry groups, such as the Vehicle Safety Communications 3 (VSC3) Consortium, operating under the Crash Avoidance Metrics Partnership (CAMP) Cooperative Agreement with the US DOT.

4 Certificate Management

Various certificate management schemes have been proposed to secure V2V communications and preserve privacy in vehicle networks. Two general categories are shared certificates and unique certificates. Traditional certificate management schemes commonly used for Internet-based communications are not suitable for vehicle communications. First, they exhibit poor privacy properties, both in terms of anonymity and linkability. Second, X.509 certificates tend to be bulky for limited bandwidth and latency sensitive communications in vehicle networks.

Securing V2V communications requires a certificate management system that supports both message authentication and privacy preservation. Implicit in this requirement is the ability to efficiently detect and effectively neutralize misbehaving vehicles. Together, these requirements create a complex problem of balancing the level of privacy and system robustness, which is defined in terms of the number of vehicles affected by certificate revocation and the ability to remove bad actors.

At one extreme, vehicle privacy is best preserved in a system where every vehicle uses the same certificate to sign V2V messages as illustrated in  Fig. 48.5. In this case, each vehicle looks like any other. However, should one or more vehicles misbehave and cause this certificate to be revoked, all vehicles in the system will require certificate replacement, and this results in significant “collateral damage” to innocent vehicles. This “one affects many” problem is an inherent attribute of shared certificate schemes where a threat to a certificate held by one vehicle is amplified by the sharing of certificates and affects a large number of vehicles. In addition, there is great difficulty in removing the vehicle(s) responsible for the revocation action in shared certificate systems.

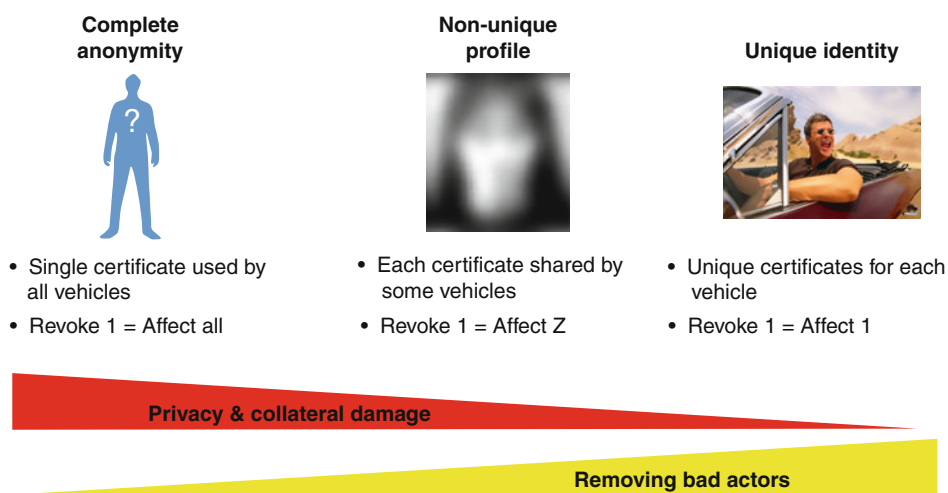


Fig. 48.5
Trade-off between privacy and system robustness

At the other extreme, the optimal system for removing misbehaving vehicles is to assign unique certificates to each vehicle. Misbehavior can be traced to a particular vehicle, and that vehicle can be removed from the system without affecting others. However, the level of privacy in this system is much lower if certain technical and operational precautions are not taken.

To better understand and contrast the benefits and trade-offs between privacy and system robustness imposed by certificate management schemes, we will examine a combinatorial certificate management scheme and short-lived, unlinked certificate management scheme.

4.1 Combinatorial Certificate Management

A popular example of a shared certificate scheme is the combinatorial certificate scheme proposed by Tengler et al. (2007a). In this method, each certificate is assigned to more than one vehicle, and each vehicle is assigned multiple shared certificates. The combinatorial certificate management scheme falls in between the two extremes previously described. A signed message can be attributed to a subset of the population, but not an individual. Similarly, certificate revocation affects more than one individual, but not the entire population. Bad actors can be narrowed to a subset of the population where additional measures and filters need to be applied. The combinatorial certificate management scheme was used in the IntelliDrive Proof of Concept Trial and is summarized below.

The Certificate Authority (CA) creates a pool of N uniformly and independently distributed triples, where each triple contains a public key, a secret key for a digital

signature scheme, and an associated certificate (i.e., a digital signature of the previous public key obtained using a master pair of public and secret keys that are associated with the CA). For privacy, N is significantly less than the total number of vehicles in the system (V), i.e., $V \gg N$. N is typically in the range of 10,000–30,000. Each vehicle will be given a small number of certificates (n) by the CA that are randomly and independently chosen from the shared pool, where $0 < n \ll N$. Certificates are not unique to any particular vehicle. Parameter n is typically in the range of 5–20.

Each vehicle randomly selects a certificate from its cache of n upon entering an RSE zone and uses it to sign V2V messages. Other triggers to select a different certificate may include expiration of a timer or using a new certificate for each V2V message. The latter technique is discouraged because an eavesdropper can very quickly learn all the certificates assigned to a particular vehicle. Knowledge of a vehicle's set of n certificates represents a unique pseudonym for the vehicle and compromises privacy. For example, the probability of choosing the same five certificates from a pool of 10,000 certificates is 1.2×10^{-18} . A novel method to select a certificate is described by van den Berg et al. (2009) where a certificate already in use in an RSE zone is chosen by an entering vehicle if it also possesses it to inconspicuously blend into the environment.

Each certificate has a lifetime set by its expiration date, after which it must be replaced. A certificate may also be revoked by the CA during its lifetime, for example, due to the detection of malicious activity from other vehicles that share the same certificate. The CA periodically publishes a certificate revocation list (CRL). Any communication signed using a certificate on the CRL will be disregarded by the vehicle population. If a vehicle holds a certificate on the CRL, it must contact the CA for a replacement. Upon receiving a re-keying request, the CA checks if the vehicle has exceeded its re-keying quota (b). If yes, the request is denied and the vehicle must submit to a manual re-keying procedure, such as physical inspection at a dealer or repair shop. Otherwise, the CA increments a re-key counter for the vehicle, selects a new certificate with a random and independent distribution from the shared pool, designates this certificate to replace the revoked certificate in the scheme structure, and sends the certificate to the vehicle. The re-key quota b is typically in the range of 5–20.

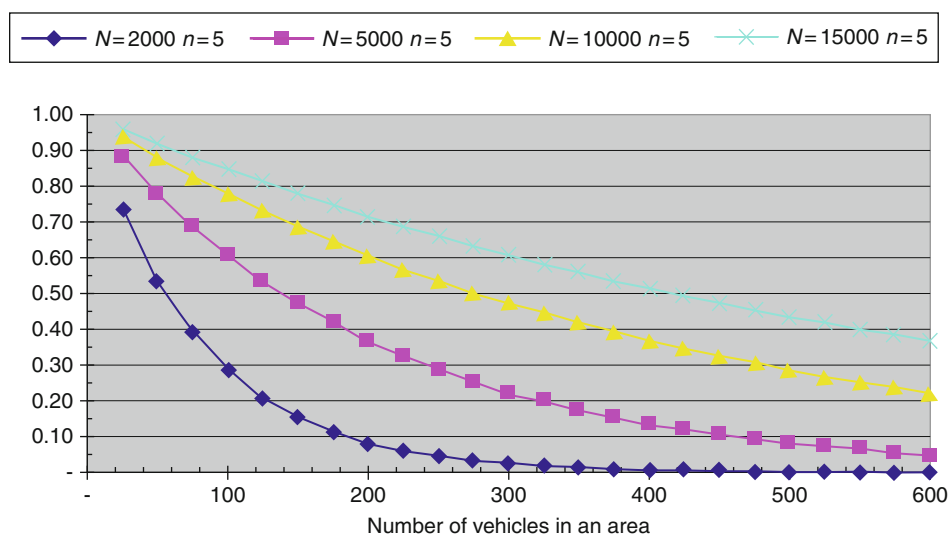
4.1.1 Privacy

In the combinatorial certificate management scheme, privacy is achieved by sharing certificates with a subset of the vehicle population. The chance that a vehicle has any particular certificate is n/N , so that $(n/N)V$ vehicles are expected to share each certificate. On a system-wide basis, each certificate will be shared by a large number of vehicles and any given transmission in a global sense cannot be associated with a particular vehicle with certainty.

Because vehicles tend to be used more for local travel than long distance travel, the probability that a vehicle will show up at any location in the system is not uniform. Consequently, although a vehicle may share a certificate with $(n/N)V$ vehicles, these

vehicles could be geographically scattered in a manner such that only a few ever cross paths. A more strenuous measure of a vehicle's privacy is the probability that all of vehicle's certificates are unique among the vehicles in a given area, such as an RSE zone. In this situation, there is one and only one vehicle using a particular certificate, and should an observer come to know the certificates for a particular vehicle, messages can be attributed to this vehicle and it could possibly be tracked. The probability that a vehicle has no certificates in common with v other units in an area is given by the Binomial probability function: $B(0, vn, n/N)$, i.e., zero successes in vn trials with success probability n/N . ● Figure 48.6 shows this probability as function of the number of vehicles in an area and the size of the certificate pool N .

As shown in ● Fig. 48.6, there is a high probability that all of a vehicle's certificates are unique in a given area, unless there are a relatively large number of other vehicles present. For the baseline system with $N = 10,000$ and $n = 5$, there is an 88% probability that a vehicle will share no certificates with its neighbors in an area with 50 other vehicles. At an intersection where there might be 25 or fewer vehicles, the probability of having all unique certificates is over 90%. The situation improves if the certificate pool is reduced to 2,000, in which case the probability of a vehicle having all its certificates unique in an area with 50 other vehicles drops to 53%. But lowering the certificate pool increases the probability that a vehicle will be affected by certificate revocation and, subsequently, shortens the period of time before it exhausts its re-key quota as described in the next section.



■ Fig. 48.6

Probability all vehicle certificates are unique in an area

4.1.2 Effect of Certificate Revocation

Because certificates are shared among many vehicles in the combinatorial certificate management scheme, a compromise of one certificate or vehicle will affect many vehicles. In a system where $n = 5$, $N = 10,000$, and $V = 200,000,000$, the number of vehicles sharing the same certificate is 100,000. If just one certificate is revoked because of a single misbehaving vehicle, approximately 100,000 vehicles are impacted.

All vehicles in possession of a revoked certificate will need to request a replacement and use up one of its allowed key replacements. Each vehicle in the system will eventually be affected by ongoing certificate revocation. If an innocent vehicle becomes the unfortunate victim of sharing a large number of certificates that have been revoked, the vehicle may trigger the enforcement of a re-keying quota. Once a vehicle has reached its re-keying quota, the vehicle is locked out, i.e., not allowed to re-key, and must be reinstated by a process that may include an investigation of the vehicle. The interval before an innocent vehicle reaches its re-keying quota is called the quota lifetime. Consider a vehicle that has been just introduced into the system with n certificates and its re-key count is set to zero. The probability of this vehicle's certificates being revoked within the first t months given m bad units per month is approximately:

$$F(t; N, n, m, b) = \left(1 - \left(1 - \frac{n}{N}\right)^{mt}\right)^b \quad (48.1)$$

The probability that a new vehicle is completely revoked in month t is given by:

$$f(t; N, n, m, b) = F(t; N, n, m, b) - F(t-1; N, n, m, b) \quad (48.2)$$

► **Equation 48.2** is the probability density function for the quota lifetime. The expected quota lifetime for a new vehicle can then be calculated by:

$$E(N, n, m, b) = \sum_t t f(t; N, n, m, b) \quad (48.3)$$

► **Figure 48.7** plots the expected quota lifetime, using baseline system parameters of $N = 10,000$ and $n = 5$ as function of the monthly bad unit rate and re-key quota b .

A first observation is that the quota lifetime for the baseline system parameters is just a few years even with a relatively small number of active bad units per month. ► **Figure 48.7** shows that the monthly bad unit rate in the baseline system can be no more than about 300 bad units per month with $b = 20$ to maintain an average quota lifetime of just 2 years. A rate of 300 bad units/month in a system of 200,000,000 vehicles means that the monthly bad unit rate cannot exceed 0.0000015 ($300/200,000,000$) or just 1.5 bad units/month per one million vehicles in the system. At this rate, a typical vehicle, whose lifetime is 12 years, will require three or more unnecessary repair visits because it was an innocent holder of shared certificates that have been misused and revoked. A second observation is that increasing the re-key quota from $b = 5$ to $b = 20$ improves the quota lifetime by less than a factor of two, but this has the adverse affect extending the period of time that a misbehaving vehicle will be allowed to remain in the system.

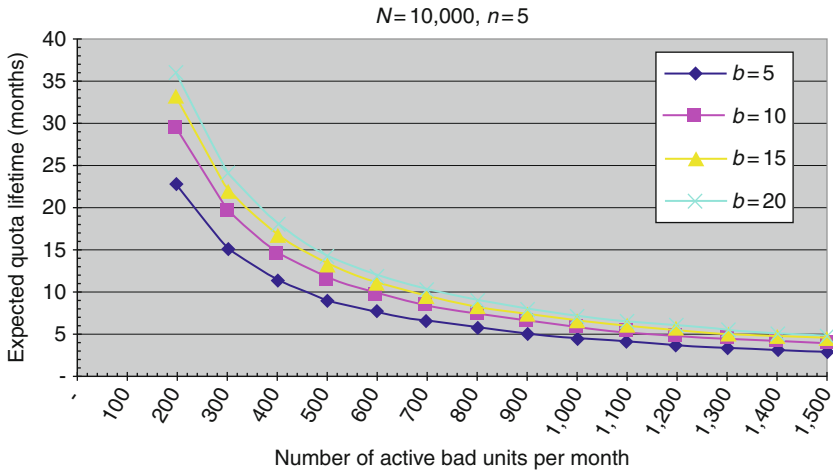


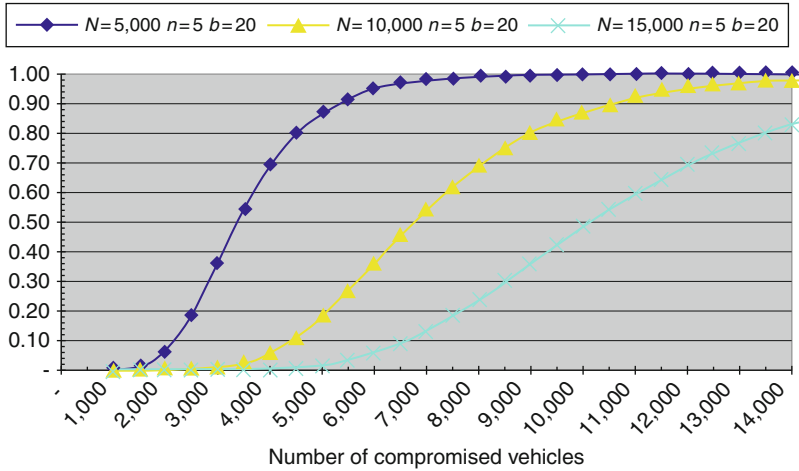
Fig. 48.7
Expected quota lifetime versus bad unit rate

The actual rate of intrusion, i.e., rate of bad units in a vehicle network, is not known and depends upon many variables. At least one such variable is the use of multiple OBE designs produced by different suppliers. A significant vulnerability in a design that affects one or more vehicle model lines could result in a sudden spike in compromised vehicles, known as a large-scale attack. One or more attackers can compromise vulnerable vehicles to either extract the certificate material or install a malicious code that uses the certificates and propagates itself. Certificate material could also be distributed through the Internet to facilitate a widespread attack.

As shown in [Fig. 48.8](#), the effect of a large-scale attack is that a large number of innocent vehicles will be locked out of the system in short time. For example, an attack that compromises 5,000 vehicles in a short time will lock out about 20% or 40 million vehicles. The recovery from such an attack could be lengthy in time and costly.

The sharing of certificates in the combinatorial certificate management scheme, while providing the benefit of anonymity, also makes it difficult to identify misbehaving vehicles. A certificate associated with a malicious message cannot be immediately traced to any particular vehicle. Using the earlier example, the misbehaving vehicle can be any one of 100,000 vehicles. Consequently, the process of identifying an intruder in the combinatorial scheme is complex. Some proposed methods use an iterative analysis that stepwise narrows the set of potential vehicles with each re-keying. Others statistically analyze the re-keying counts per vehicle and waits for the misbehaving vehicle to sufficiently stand out to meet a target error rate.

For more information and analysis on the Combinatorial Certificate Management scheme (see White et al. [2009](#)). Information about refinements to the combinatorial certificate scheme can be found in Telcordia Technologies ([2007](#)).



■ Fig. 48.8
Fraction of vehicles locked out in large-scale attack

4.2 Short-Lived, Unlinked Certificates

Short-lived, unlinked certificates (Pietrowicz et al. 2010) is an alternative certificate management scheme that avoids the “one affects many” problem of shared certificate schemes and provides a straightforward method to disable misbehaving vehicles. The basic principles of the short-lived, unlinked certificate scheme are:

1. Certificate authority partitioning is used to separate certificate authorization and assignment functions. The Authorizing Certificate Authority has no knowledge of the vehicle certificates. The Assigning Certificate Authority has no knowledge of the vehicle identity.
2. Each vehicle is assigned a large number of unique, unlinked certificates ($n \sim 100\text{--}1,000$) by an Assigning Certificate Authority.
3. Certificates are assigned in such way that the Assigning Certificate Authority does not know the complete set of certificates held by any vehicle.
4. Certificates are not ordinarily revoked. Instead, certificates are short-lived, and misbehaving vehicles are identified and removed from the system during the certificate replacement process.

The short-lived, unlinked certificate scheme overcomes the “one affects many” problem by assigning unique certificates to each vehicle. By not sharing certificates, any misuse that leads to the invalidation of a certificate affects only the compromised vehicle. The threat of an attacker learning all of the unique certificates for a particular vehicle is diminished by the use of a large number of certificates. A vehicle would need to be tracked for a long period under conditions where a message can be eavesdropped on the air link and unequivocally linked to the vehicle as its source for an attacker to learn all of a vehicle’s certificates. In this respect, vehicle privacy is improved over the combinatorial certificate

scheme, where each vehicle possesses a small number of shared certificates that are essentially unique unless there are hundreds, if not thousands, of vehicles in a given area. The probability that the same certificate is selected again in the short-lived, unlinked certificate scheme is much lower than the combinatorial scheme.

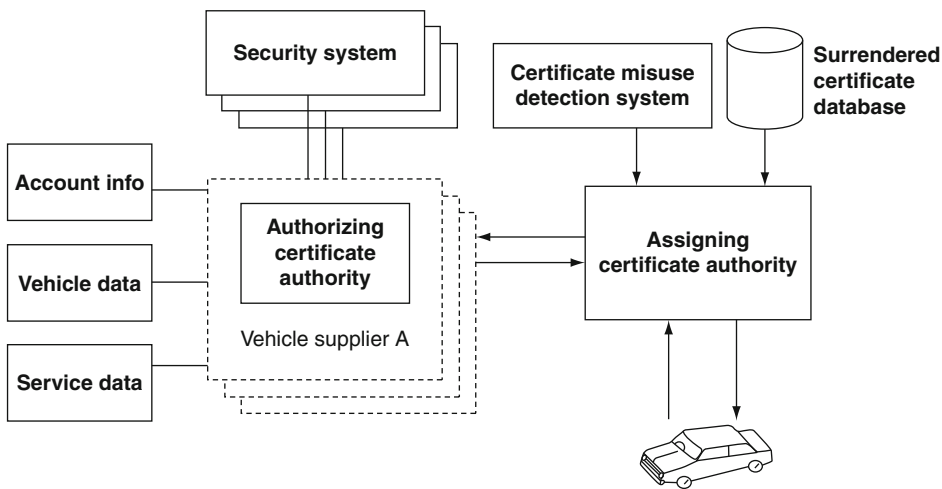
4.2.1 Privacy

The use of unique certificates in the short-lived, unlinked certificate scheme, however, does require special techniques to protect the anonymity of vehicles. In particular, no entity other than the vehicle holding certificates should know of its complete certificate set nor should any single entity be able to associate a set of certificates with a particular vehicle, even if the vehicle’s identity is not known. As previously mentioned, the vehicle network itself is in the best position to track a vehicle, especially if it is capable of associating a large number of unique certificates with a particular vehicle.

The short-lived, unlinked certificate management scheme uses the following two techniques to provide vehicle anonymity:

- 1. Certificate Authority Partitioning
- 2. Unlinked Certificate Assignment

► *Figure 48.9* illustrates the concept of certificate authority partitioning in the short-lived, unlinked certificate scheme. The principle components are the Authorizing Certificate Authority, the Assigning Certificate Authority, and the vehicle.



■ Fig. 48.9
CA partitioning in short-lived, unlinked certificate management scheme

4.2.2 Authorizing Certificate Authority Functions

The Authorizing Certificate Authority has three primary functions. One function is to issue the long-term vehicle-identifying certificate, which is unique to each vehicle. The identifying certificate is assigned during vehicle manufacture preferably through a process where the vehicle generates the long-term identifying key pair. The Authorizing Certificate Authority also makes its public key known to the vehicle when it assigns the long-term identifying certificate.

A second function of the Authorizing Certificate Authority is to authorize vehicle requests for certificates via a proxy through the Assigning Certificate Authority. A security system performs authorization checks on every request for certificates. The authorization checks may include:

- *Is the vehicle registered with the Authorizing Certificate Authority?*

This authorization check is intended to deny attackers who intentionally direct their requests to the wrong Authorizing Certificate Authority in an attempt to acquire certificates.

- *Has the vehicle been reported stolen?*

When a vehicle is reported stolen, vehicle network policies may dictate whether the vehicle is allowed to re-key.

- *Has the vehicle been recycled or salvaged?*

Although not essential, it would be beneficial for the Authorizing Certificate Authority to know whether a vehicle is no longer in service. If a vehicle has been recycled or salvaged, all attempts to re-key using its long-term identifying certificate should be denied. This would help prevent attackers from maliciously using OBE acquired from salvage yards.

- *Is the vehicle of a model whose vehicle network functions are known to have been compromised?*

Each vehicle manufacturer will develop and deploy its own version of an OBE. Some vehicle manufacturers will deploy different versions of OBE based on the class of vehicle. In addition, OBE implementations will periodically be redesigned to incorporate new features, new hardware, and cost reductions. Each OBE implementation has the potential to inadvertently introduce a serious vulnerability in the vehicle network. Some vulnerabilities may be patched by a software update. Others may require replacement of the OBE. Once an attacker identifies a particular model of OBE with an exploitable weakness, the vehicle network requires a mechanism to deny re-keying to affected vehicles until they have been repaired.

- *Does the vehicle contain an OBE that needs to be replaced or is no longer supported in the vehicle network?*

The vehicle network is intended to be a long-term component of the national highway infrastructure. Improvements and other changes to the vehicle network may occur after its deployment. At some point, a system modification may be introduced that, for instance, provides positive safety benefits, but is not compatible with certain

vintages of OBE. This authorization check provides a means for the vehicle network to exclude certain vehicles as active participants until they are upgraded.

- *Has the vehicle been locked out?*

In the short-lived, unlinked certificate scheme, vehicles that have been denied a re-keying request, for instance, because they are associated with certificates that have been misused, can be locked out of the system and denied all future re-keying requests until they have passed a more stringent test, such as a vehicle inspection.

- *Has the vehicle been re-keying excessively?*

This authorization check is designed to detect and deny re-keying requests if a vehicle re-keys more often than expected. It can be used to detect attacks where, for instance, expired certificates from salvage vehicles are obtained and used by an attacker to increase the size of their certificate pool. By increasing the number of certificates at their disposal, attackers may spread their attack across a large number of certificates and potentially make their activity fall below the detection threshold of the vehicle network. The Authorizing Certificate Authority may maintain a count of all re-keying requests and a sliding window of the number of re-keying requests that were authorized. If a vehicle re-keys more than an expected amount, the vehicle can be considered suspect and denied. The window duration and re-keying quota should take into consideration the certificate lifetime, number of certificates per vehicle, and number of requests needed to replace a vehicle's complete set of certificates once.

- *Is the vehicle still in the initialization stage?*

This authorization check is designed to detect if a vehicle has not yet completed initialization and has received its first set of n certificates. The Authorizing Certificate Authority temporarily keeps track of the number of certificates provided to each vehicle. When the vehicle has received n certificates, a flag is set to indicate that the vehicle is initialized. The Authorizing Certificate Authority informs the Assigning Certificate Authority about whether the vehicle has past the initialization stage and should be able to surrender expired certificates.

- *Is the vehicle requesting fewer keys over time?*

This authorization check is designed to detect a "Sequential Certificate Use" attack that is described in [Sect. 4.2.6](#). The Authorizing Certificate Authority observes the number of certificates requested by a vehicle over time and detects vehicles that are consistently requesting few total certificates within a certificate interval. This security check can be done off-line and does not need to be performed during a certificate request process.

A third function of the Authorizing Certificate Authority is to optionally store other identifying information about each vehicle, such as its vehicle identification number, build options, configuration, software versions, owner info, and potentially subscriber account information for original equipment manufacturer (OEM) services. As shown in [Fig. 48.9](#), the Authorizing Certificate Authority is supported by several supporting functions, such as account info, vehicle data, and service data sources. Because the Authorizing Certificate Authority contains vehicle-identifying information, it is assumed to be privately held by vehicle manufacturers.

4.2.3 Assigning Certificate Authority Functions

The Assigning Certificate Authority in the short-lived, unlinked certificate scheme is tasked with processing vehicle requests for certificate replacement and publishing emergency CRLs to vehicles. It is supported by a Certificate Misuse Detection System, which provides the Assigning Certificate Authority with a list of certificates that have been misused, and a Surrendered Certificate Database, which keeps track of surrendered certificates. The Assigning Certificate Authority performs a series of validation checks on each certificate request. If the validation checks are successful and the request is authorized by the Authorizing Certificate Authority, the vehicle is assigned one or more certificates with a short-lived expiration, which is on the order of several weeks.

It is essential to prevent the Assigning Certificate Authority from building certificate associations, even if it does not know to which vehicle a set of unique certificates belongs. The reason is to prevent an adversary from eavesdropping on a target vehicle, identifying at least one of its certificates, and then making use of the vehicle network to retrieve the entire set of certificates that belongs to the target vehicle. If the vehicle network knows the set of unique certificates for a vehicle, it can potentially implement roadside equipment (RSE) probes and promiscuously collect, filter, and monitor transmissions of a particular vehicle.

The short-lived, unlinked certificate scheme prevents that Assigning Certificate Authority from building a complete set of certificates for a vehicle by using an unlinked certificate assignment method, where certificates are not all assigned to a vehicle in a single request. Instead, the certificates are assigned over the course of multiple transactions that are distributed over a period of time to take advantage of the mixing of requests from a large number of vehicles. The requests are not periodic, but, instead, they are made at random intervals within a predefined window to prevent the Assigning Certificate Authority from inferring certificate associations through a temporal analysis of the requests.

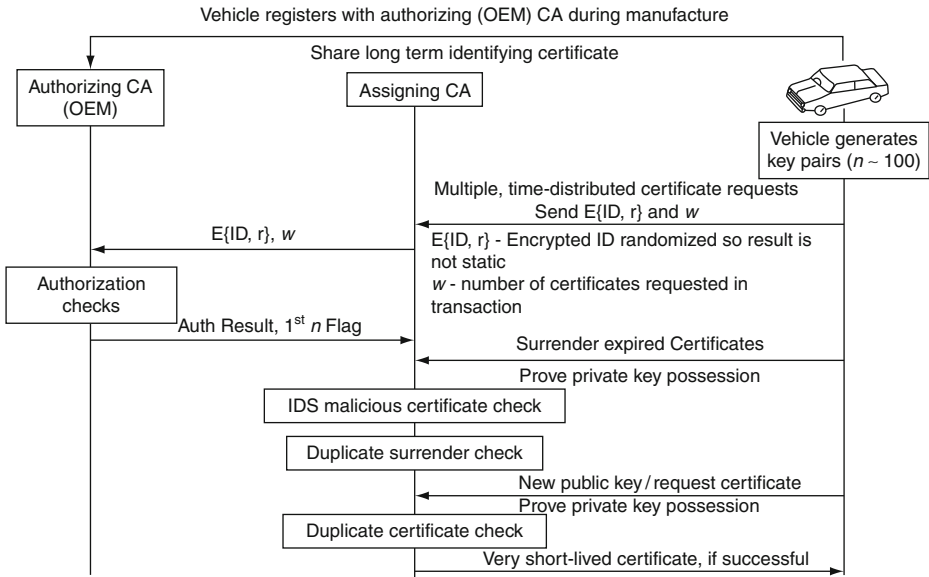
The Assigning Certificate Authority may on rare occasions publish an emergency CRL to revoke certificates that are causing significant disruption to the system. Unless there is a particularly disruptive attack, the network would rely upon certificate expiration to disable the misbehaving vehicle. Since certificate life times are short, the emergency CRL quickly decreases in length.

Optionally, the Assigning Certificate Authority can treat V2V and V2I communication separately and publish CRLs for RSEs, which have a more reliable and higher bandwidth network connection and are less resource constrained.

4.2.4 Certificate Request Transaction Flow

The certificate request transaction is central to the short-lived, unlinked certificate scheme. It is the mechanism that:

1. Maintains the separation of vehicle identity and certificate assignment
2. Prevents the association of certificates, even for an unidentified vehicle
3. Identifies and removes intruders from the vehicle network



■ Fig. 48.10
Short-lived, unlinkable certificate request transaction flow

A critical aspect of preventing the linking of requests is to eliminate the use of static identifiers between requests. In the short-lived, unlinkable certificate method, the Authorizing Certificate Authority needs to know the long-term identity of the vehicle. To satisfy this need and not reveal the vehicle identity or a static identifier associated with the vehicle to the Assigning Certificate Authority, the vehicle passes its encrypted long-term identifier and the number of certificates requested (w) in the certificate request. The long-term identifier contains a timestamp, nonce, or other temporal value so that the encrypted result is not static and cannot be used by the Assigning Certificate Authority to link certificate requests. The long-term identifier can be the vehicle's long-term identifying certificate signed with its long-term identifying private key and encrypted using the public key of the Authorizing Certificate Authority.

➤ **Figure 48.10** illustrates the transaction flow for the certificate replacement process. It assumes that each vehicle has previously registered with an Authorizing Certificate Authority, which has issued a long-term identifying certificate to the vehicle. Registration preferably takes place during vehicle manufacture. Although the vehicle can be loaded with short-term certificates during registration, ➤ **Fig. 48.10** advantageously assumes the vehicle acquires its first set of certificates using the same process that is used for certificate replacement.

A detailed description of the certificate request transaction flow is presented below:


1. *Vehicle Generates Public/Private Key Pairs:* Before the vehicle initiates the certificate replacement transaction, it generates a set of unique public/private key pairs for which it will seek certificates from the Assigning Certificate Authority. Because the

vehicle generates the public/private key pairs as opposed to being assigned key pairs by a certificate authority, knowledge of the private keys is advantageously held by one only entity in the vehicle network, i.e., the vehicle that generated and uses them.

2. *Multiple, Time-Distributed Certificate Requests:* Upon creating the public/private key pairs, the vehicle launches a series of requests to obtain certificates from the Assigning Certificate Authority. Unlike previous methods, the vehicle does not acquire all of its certificates in a single request. Instead, multiple requests that are distributed over time take advantage of the natural anonymity provided by the statistical mixing of requests from many vehicles. In each request, the vehicle attempts to obtain w certificates, where w is randomly selected by the vehicle. The maximum value of w (W) can be a system parameter to help adjust the volume of certificate replacement requests. The Assigning Certificate Authority, for instance, can set parameter W , but the vehicle can select to request any number of certificates from 1 through W . In this manner, the vehicle network has a limited ability to force associations among certificates through knowledge of what certificates were provided in a single request.

The certificate requests are distributed over a spreading period of time, $t_{\text{spreading}}$, and the intervals between requests are randomized such that the Assigning Certificate Authority cannot successfully perform a temporal analysis on the requests and link requests from a vehicle. The time period $t_{\text{spreading}}$ may be on the order of the certificate lifetimes. However, detailed analysis is needed to judiciously choose the range given the desire to quickly remove malefactors, manage re-keying volume, and maintain the unlinkability among certificate requests.

3. *Randomized, Encrypted Identity:* To acquire certificates, the vehicle must present its long-term identity. However, presenting its identity in the clear would provide the Assigning Certificate Authority with a static identifier with which to link the multiple, time-distributed requests. Simply encrypting the vehicle identity is also not sufficient because the encrypted result would be static and it could be used as a vehicle identifier to link certificate requests. Instead, the short-lived, unlinked certificate scheme requires the vehicle to encrypt its identity with a temporal value, such as a timestamp or random nonce, such that the encrypted result varies. The vehicle uses the public key of the Authorizing Certificate Authority to asymmetrically encrypt the combination of its identity and temporal value. In one implementation, the binary value of the vehicle's identity can be interleaved with a random nonce to sufficiently randomize the encrypted result.
4. *Request Authorization:* Upon receiving a certificate request, the Assigning Certificate Authority queries the Authorizing Certificate Authority for approval. It passes the randomized, encrypted vehicle identity and the number of certificates, w , which are being requested in the transaction. The Authorizing Certificate Authority decrypts the randomized, encrypted vehicle identity using its private key and performs a series of authorization checks as previously described. The Authorizing Certificate Authority returns an authorization result to the Assigning Certificate Authority along with an indication of whether the vehicle has been fully initialized with its first set of certificates.

5. *Surrender Expired Certificates*: Assuming the authorization result is positive, the Assigning Certificate Authority signals the vehicle to surrender its expired certificates. The short-lived, unlinked certificate scheme requires each vehicle to present its expired certificates and prove ownership of those certificates. For each certificate that is requested, the vehicle must present an expired certificate. It must prove ownership of the private key by, for instance, signing the certificate with the corresponding private key. The only exception is the case where a vehicle is still initializing and has not yet received its first n certificates. In this case, the Assigning Certificate Authority has been informed by the Authorizing Certificate Authority that the vehicle is still initializing and the Assigning Certificate Authority can grant it an exemption from surrendering the proper number of expired certificates.
6. *Malicious Certificate Check*: The Assigning Certificate Authority receives a list of certificates that its Certificate Misuse Detection System has identified as being compromised or associated with malicious behavior. The Assigning Certificate Authority compares each of the expired certificates received from the vehicle against this list. If any of the surrendered certificates are on the misused certificate list, re-keying is denied and, preferably, the Authorizing Certificate Authority is informed and makes record that the vehicle is locked out and needs to be inspected.
7. *Multiple Surrenders of the Same Certificate*: Because the Assigning Certificate Authority does not know the identity of the vehicle making a certificate request or the set of certificates that a vehicle possesses, the Assigning Certificate Authority must protect against an attack where a vehicle attempts to surrender the same expired certificate in multiple requests. Although other solutions may be possible, the most obvious solution is to maintain a database of surrendered certificates as shown in  Fig. 48.10. Every time a certificate is surrendered, the database is queried to determine if it has previously been surrendered. If not, the certificate is added to the database and the re-keying process continues. Otherwise, the re-keying process is halted and the vehicle may be locked out. The database can be a flat file that is distributed to maximize search efficiency. Back-end processing can sort and organize the database to improve the search. It may be possible to safely delete surrendered certificates that are older than a given time period.
8. *Prove Ownership*: The Assigning Certificate Authority confirms that the vehicle making the certificate request has the private key for the certificate being requested.
9. *Duplicate Certificate Check*: The check for duplicate certificates is optional and depends largely upon the probability that two or more vehicles will generate the same public/private key pair and request a certificate in the same time interval. Certificates need to remain unique so that a misused certificate can be traced to a single vehicle. The Assigning Certificate Authority can perform this check by maintaining a list of all certificates that it issued over its current certificate interval. For instance, if the Assigning Certificate Authority uses a granularity of 1 day in its certificate expiration date, the Assigning Certificate Authority would only need to maintain a list of certificates that it issued in the current day. To perform the duplicate certificate check, the Assigning Certificate Authority would generate a new certificate

and compare it against the list of certificates issued in the current certificate interval. If a duplicate is found, the vehicle is requested to generate an alternate public/private key pair. To prevent disclosing that a key pair generated by a vehicle has a duplicate, the Assigning Certificate Authority can randomly request vehicles to generate alternate public/private key pairs. If the duplicate check is not performed and certificate duplication occurs, the result is that one vehicle would lose a certificate, which is not serious given that n is approximately 100. If the duplicate certificate is additionally used maliciously by a vehicle, both vehicles could be locked out.

10. *Assign Short-Lived Certificates:* If the vehicle has satisfied all the above Authorizing and Assigning Certificate Authority checks, the Assigning Certificate Authority generates and sends the vehicle one or more short-lived certificates. The lifetime of the certificates is on the order of weeks and can be adjusted based on the characteristics of the malicious message detection algorithms, the amount of malicious activity detected, and available certificate authority capacity.

4.2.5 Detection and Removal of a Misbehaving Vehicle

The short-lived, unlinked certificate scheme relies upon the principle of “frequently proving innocence” to identify misbehaving vehicles. The concept of “frequently proving innocence” is implemented by using short-lived certificates, which preferably have lifetimes on the order of several weeks. Misbehaving vehicles are identified and purged during the certificate replacement process. Specifically, each vehicle must surrender its expired certificates to acquire a replacement. Replacement of expired certificates is done on a one-for-one basis. By surrendering its expired certificate, the vehicle enables the Assigning Certificate Authority to determine its innocence or guilt. If the vehicle surrenders a certificate that was associated with malicious activity, the vehicle is denied and locked out. Similarly, if any vehicle attempts to surrender a certificate that was previously surrendered, it is denied and locked out as this action in itself is malicious.

The use of short-lived certificates provides a couple of additional benefits beyond frequently proving innocence. First, short-lived certificates may remove the need to publish a CRL. When expiration dates are used with certificates, the purpose of the CRL is essentially to address serious threats that cannot wait for the certificates to expire. If the expiration period is short, the delay in distributing a CRL may be on the same order as the expiration period. If the system can tolerate a short period of malicious behavior, the process can be simplified by eliminating the CRL altogether. Second, the ability to forgo publishing a CRL eliminates a major problem with unique certificate schemes, i.e., CRLs for unique certificates grow to be very large, very quickly. In the short-lived, unlinked certificate scheme, CRLs are not published to vehicles.

The certificate lifetime in the short-lived, unlinked certificate scheme can be an adjustable system parameter. To some extent, this helps to address the issue of an unknown number of attackers or rate of attack. The lifetime of a certificate can be adjusted, for instance, based on the amount of malicious activity detected in the vehicle

network. If the amount of malicious activity is low, certificate lifetimes can be increased to lower overall certificate authority load. If the amount of malicious activity increases, the certificate lifetimes can be shortened to more quickly purge misbehaving vehicles. In this manner, the system can be “throttled.” One constraint is that the certificate lifetime should not be made shorter than the time it takes the system to detect malicious behavior. Otherwise, certificates get replaced before malicious behavior can be detected.

Another potential benefit is that certificate lifetimes and other system parameters can be tailored on a regional basis without much coordination among the different regions. For instance, certificate lifetimes can be based on RSE coverage and population densities. Metropolitan areas with good RSE coverage can use shorter certificate lifetimes because vehicles frequent areas with RSE coverage and can conveniently complete certificate replacement transactions. In rural areas where RSE coverage is sparse, longer certificate lifetimes can be used to help reduce the possibility that all certificates expire before the vehicle re-enters RSE coverage. The primary coordination that is required among regions is the sharing of the misused certificate list and the surrendered certificate database.

4.2.6 Weaknesses

The short-lived, unlinked certificate scheme has some weaknesses. Notably, it is vulnerable to a sequential certificate use attack and requires massive certificate renewals for a large vehicle network.

The sequential certificate attack attempts to maximize the time that a misbehaving vehicle can continue to operate with valid certificates. To implement this attack, an attacker uses one or only a small number of certificates to attack and stores the remainder of the certificates. When the certificates expire, the attacker requests replacements, but surrenders only the expired stored certificates. The attacker then repeats the process. The result is that the attacker’s cache of certificates constantly decreases, but its ability to operate becomes a multiple of the certificate lifetime.

The short-lived, unlinked certificate scheme must also protect an attack where a certificate not associated with malicious behavior is surrendered more than once either by the same vehicle or multiple vehicles to “prove innocence” and acquire a replacement certificate. One method is to maintain a database of all certificates that have been surrendered by all vehicles. A concern with this approach is the size and practicality of such a database. To estimate the size of the Surrendered Certificate Database, assume the following system parameters:

- Each vehicle has 100 certificates, i.e., $n = 100$.
- Each vehicle uses 10 batches to acquire 100 certificates.
- Each certificate expires in 2 weeks.
- There are 200 million vehicles in the vehicle network.
- Certificates in the Surrendered Certificate Database are deleted after two years.

Based on the above system parameters, the Surrendered Certificate Database would need to make an entry for:

$$2 * (200 \times 10^6 * 100 * 52/2) = 1.04 \times 10^{12} \text{certificates} \quad (48.4)$$

If there are 100 Surrendered Certificate Databases, each database would contain 10 billion certificate entries, which makes for a very large database.

The number of certificate replacement transactions per year is:

$$10 * (200 \times 10^6 * 100 * 52/2) = 5.2 \times 10^{12} \text{certificate replacement transactions} \quad (48.5)$$

If there are 100 Surrendered Certificate Databases, each database would need to process about 1,650 certificate replacement transactions/s. As a point of reference, an individual Domain Name Service (DNS) server proposed for the .org top-level domain has a capacity in the range of 5,000 queries/s (<http://www.icann.org/tlds/org/questions-to-applicants-11.htm>).

For more information and analysis on the short-lived, unlinked certificate management scheme (see Pietrowicz et al. 2010).

5 IntelliDrive Security Architecture

The basic components of the IntelliDrive security architecture are described below:

1. Uses a PKI system that supports both shared and unique IEEE 1609.2 certificates. Shared certificates are used for applications where privacy is required. Unique certificates are unused for applications where the user/entity needs to be identified.
2. The combinatorial certificate management scheme where $N = 10,000$ and $n = 5$ is used as the shared certificate management scheme.
3. RSEs are issued unique certificates. OBE are issued shared and unique certificates, depending upon application.
4. Safety applications use IEEE Wireless Access in Vehicular Environments (WAVE) Short Messages signed by a shared certificate in the OBE and by a unique certificate when broadcast by an RSE.
5. The Probe data collection application is secured using a new secure session protocol called Vehicular Datagram Transport Layer Security (VDTLS) (Pietrowicz et al. 2008) that protects anonymity by using shared certificates to negotiate a UDP/IP secure session.
6. The Navigation application is secured using a new protocol called Vehicular Host Identity Protocol (VHIP) that protects anonymity to negotiate a TCP/IP secure session.
7. Toll and Parking Payment applications are secured by signing and asymmetrically encrypting UDP/IP and TCP/IP packets, respectively.
8. Certificate Management applications are secured by signing and asymmetrically encrypting TCP/IP packets.
9. RSEs are issued certificates with geographic attributes. RSEs evaluate their current GPS location against the location attribute in the certificate issued by the IntelliDrive CA before broadcasting.

10. OBE and RSE use geographic scope attributes in certificates to filter received messages.
11. Timestamps and cryptographic nonces are used to prevent replay attacks over the air.

► **Table 48.1** summarizes the protocols and security for the IntelliDrive applications.

For more information about the IntelliDrive Proof of Concept Trial (see <http://www.intellicdriveusa.org/library/research-reports/technical/dsrc-poc.php>).

■ **Table 48.1**

Summary of IntelliDrive security by application

Application	Transport protocol	Security	Network security endpoint
Positioning service	WAVE short message (WSM) on control channel	Signed WSM	RSE proxy
Signage	WSM on control channel	Signed WSM	RSE proxy
	WSM on service channel		
Geographic intersection description (GID)	WSM on control channel	Signed WSM	RSE proxy
CICAS signal phase & timing	WSM on service channel		
HeartBeat	WSM on control channel	Signed WSM	None
Probe data collection (w/o VDTLS)	UDP/IP on service channel	None	RSE proxy
Probe data collection (w/VDTLS)	UDP/IP on service channel	Secure anonymous session using vehicular datagram transport layer security (VDTLS)	RSE proxy
Local e-payment – parking	TCP/IP on service channel	Signed and encrypted (asymmetric)	RSE proxy/local transaction processor
Local e-payment – toll	UDP/IP on service channel	Signed and encrypted (asymmetric)	RSE proxy/local transaction processor
Network e-payment	TCP/IP on service channel	Vehicular host identity protocol (VHIP)	Network users payment (NUP)
Navigation	HTTP/TCP on service channel	VHIP	Transaction service manager (TSM)
Traveler information	WSM on service channel	Signed WSM	RSE proxy
Certificate manager	HTTP/TCP on service channel	Signed and encrypted (asymmetric)	Certificate authority

6 Intrusion Detection in Vehicle Networks

Intrusion detection for vehicle networks is a three-step process as illustrated in [Fig. 48.11](#). The process starts with an accurate and reliable capability to identify malicious communications. Once a malicious message is detected, the certificate that was used to generate the message is recorded, along with other information associated with the intrusion occurrence. In the second step, misused certificates are analyzed to narrow the impact zone of the intrusion and attempt to distinguish the source of the malicious messages from the vehicle population. At the appropriate time, the vehicle certificate authority takes action to remove bad actor from the vehicle network by revoking the misused certificate and denying re-keying requests.

As shown in [Fig. 48.12](#), the primary components in vehicle networks that can actively participate in intrusion detection for vehicle communications are:

- Vehicles equipped with OBE
- Roadside equipment
- Network-based intrusion detection analysis systems
- The 1609.2 CA that assigns and replenishes vehicle certificates and generates CRLs
- Back-end application servers

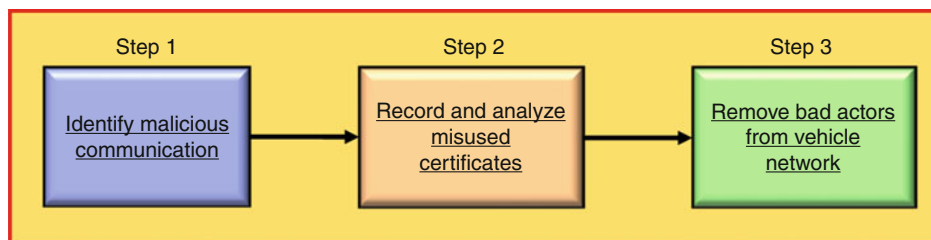
The rest of this section discusses intrusion detection techniques that can be applied in different portions of the vehicle network.

6.1 Intrusion Detection by OBEs

OBEs are a primary target for threat agents. Various methods have been previously described for how an OBE might be compromised and used to generate malicious messages. Some of these methods include altering sensor data input to the OBE, physically tampering with the OBE, and inserting malicious code within an OBE application. In addition, vehicle communications can be manipulated by a threat agent equipped with a DSRC-capable device, such as a laptop, which can generate and receive DSRC messages.

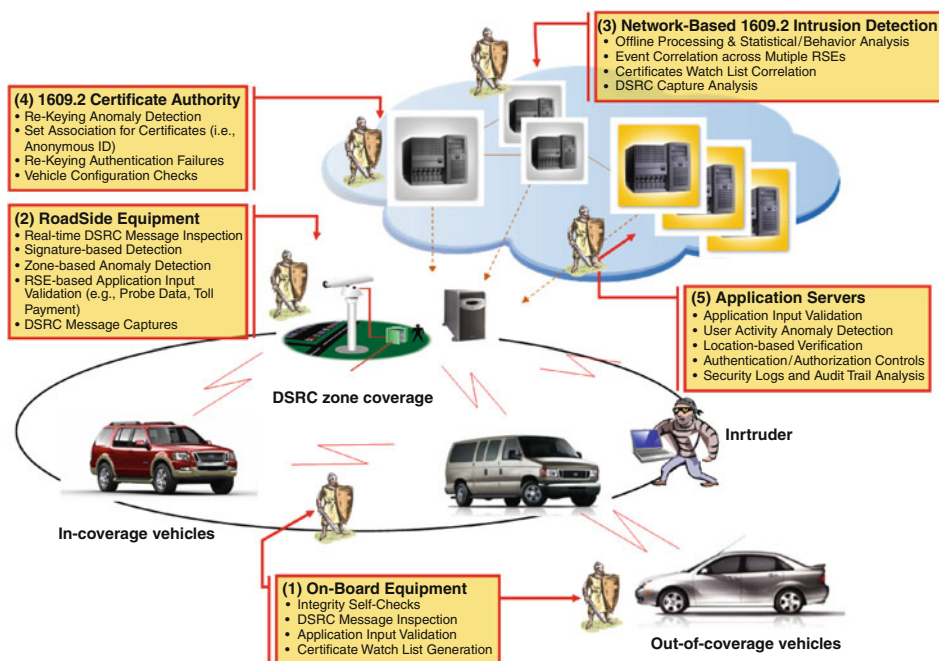
OBEs will require some intrusion detection and self-protection capabilities. When vehicles leave an RSE zone, they enter an unmanaged environment. They continue to communicate with each other but without the trusted vehicle infrastructure having visibility over the messages exchanged among vehicles. In this environment, the only means of detecting intruder activity is through the OBE.

The type of intrusion detection that can be incorporated on an OBE needs to take into account that it has limited resources and processing power. It is unlikely that OBEs will have the capacity to execute computationally intensive intrusion detection algorithms. OBE intrusion detection capabilities need to support but not interfere with its normal operation. It cannot delay, for instance, the processing of heartbeat messages to the point where they are no longer relevant.



■ Fig. 48.11

Intrusion detection process



■ Fig. 48.12

Intrusion detection components for vehicle networks

In addition, the OBE is at high risk of attack. While this may argue for very sophisticated capabilities, there is only a limited amount of trust that can be placed in declarations made by any individual OBE. There is always the possibility that an intelligent attacker could compromise the intrusion detection capabilities and begin to use them to its advantage.

Given these constraints, there are at least four forms of intrusion detection that can be implemented within OBEs. They include:

- OBE Integrity Self-Checks
- DSRC Message Inspection

- Application Input Validation
- Certification Watch List Generation/Reporting

The OBE can perform one or more integrity self-checks on both its hardware and software to look for evidence of tampering or intrusion. One form of integrity check can determine if the OBE is still connected to the vehicle in which it was initially installed. This could be achieved by acquiring unique vehicle descriptors, hashing the set, and comparing the result against a stored value. Another form of integrity check is to verify the condition of various internal hardware components. A memory chip that has been erased or lacks a particular signature could indicate potential intrusion. Likewise, OBE firmware, operating system files, and applications can be checked via hash signatures. Some hardened versions of components can actually detect the manipulation of input signals, component power, clock frequency, etc., and take actions to disable themselves. Integrity checks are also possible within the OBE application environment. For instance, some Open Services Gateway Initiative (OSGi) frameworks incorporate mechanisms to detect runaway or non-returning applications and disable those applications upon the next restart.

OBE integrity self-checks can be used to validate sensor data. Electronic control systems in vehicles commonly include capabilities to detect malfunctioning sensors and either ignore or substitute data to keep the vehicle operational. OBEs can similarly check sensor data for signs that the sensor is being manipulated. For instance, a change in GPS coordinates that is physically not possible for the vehicle to achieve over an interval of time might be indicative of intrusion. Sensor data that indicates vehicle braking but which conflicts with vehicle speed and accelerometer information could be indicative of sensor manipulation in an attempt to generate bogus emergency brake light messages.

These simple examples demonstrate that much information is actually available to the OBE to perform integrity self-checks. However, making a determination as to what is normal operation, a vehicle or component fault, or evidence of intruder activity can be a difficult task. Furthermore, if an integrity self-check fails, how should an OBE respond? The answer is likely to be dependent upon the particular failure. But some possible actions, which are not exclusive of each other, include the OBE (1) shutting itself down, (2) disabling itself by erasing all or part of its memory, (3) ignoring the faulty input, but tracking its occurrence, and (4) reporting back to the vehicle network infrastructure that a particular set of anonymous certificates are potentially at risk of compromise.

It is expected that the OBE will perform standard message inspection techniques to check the taxonomy of DSRC messages. The inspection will need to be performed in real-time. Inspection techniques will be limited by performance considerations for packet delay and throughput. Many vehicle safety messages, such as for crash avoidance, are time-sensitive and occur in abrupt batches. Malformed messages that have valid signatures might be considered evidence that a misbehaving vehicle is within communications range. Signature-based techniques might also be implemented. But this requires an OBE capability to periodically update the attack signature library.

Attacks that are more intricate will likely seek to manipulate application data within a properly formatted DSRC message. General packet inspection techniques will be unable to validate application data. Each application on an OBE will need to perform some level of validation of the information it receives from other vehicles, and possibly RSEs. Even though a DSRC message has a valid signature, it could have been maliciously generated by an intruder that has gained access to valid 1609.2 private keys.

Applications will need to perform validation of input using methods such as value range checking, sequence analysis, conformance to protocol semantics, violation of physical laws, etc. If exceptions are detected, the vehicle may place the certificate associated with the message on a watch list, where upon excessive abnormal messages may cause the vehicle to ignore all messages with this certificate while within a particular zone or for a time period, particularly when the vehicle is not within DSRC zone coverage. The OBE can also periodically report its certificate watch list back to the vehicle network infrastructure, whereupon correlation and anomaly detection techniques could be used to detect certificate misuse. The use of certificate watch lists in conjunction with network-based intrusion detection is further described in [Sect. 6.3](#).

6.2 Intrusion Detection by RSEs

RSEs have a several advantages over vehicles for performing intrusion detection:

- The RSE is network element that can be trusted by the intrusion detection algorithms running in the vehicle network infrastructure. If an RSE determines that a certificate has been misused, its declaration will not be challenged by the vehicle network infrastructure to the same extent that a vehicle would be if it made a similar declaration.
- RSEs will have more capable computing platforms than OBEs and will be able to perform more complex intrusion detection algorithms. In addition to performing the packet inspection methods used by OBEs, RSEs will be able to use more signature-based methods. As part of the vehicle network infrastructure, RSEs can be periodically updated with attack signatures and can monitor for known forms of system abuse.
- RSEs can more easily apply behavior-based intrusion detection techniques. Each RSE can collect information about the vehicle traffic patterns in its zone and establish a norm for DSRC messaging, possibly as a function of traffic density, time of day, and day of week. If an RSE observes DSRC message traffic deviating from the norm, the RSE could employ additional methods to determine if certificates are being misused. One possible method is to establish a norm for the number of DSRC messages that a vehicle sends while in the zone, possibly distributed by the type of message. Substantial deviation could indicate certificate misuse.
- Some RSEs will host very specific applications, such as probe data collection and toll payment. In both cases, RSEs can perform application input validation because they

act as an endpoint for the communication with the vehicle. RSEs can, for instance, collect probe data from multiple vehicles and perform consistency checks among the various data. If one vehicle reports data that indicates major congestion in the area but other vehicles report a low traffic, data from the outlying vehicle may be considered suspect.

In the case of a toll payment application, RSEs have a significant advantage because toll payment is a service that requires each vehicle to identify itself or an account associated with the vehicle. Checks can be performed off-line to validate the account information provided. If the account is bogus or stolen, the transaction might be traced back to the OBE certificate that was used to securely conduct the transaction.

6.3 Network-Based Intrusion Detection

Whereas RSEs focus on intrusion events within their zone of coverage, network-based intrusion detection can be used in vehicle networks to monitor intrusion activity across multiple RSEs. By virtue of its broader perspective, network-based intrusion detection can collect information from multiple RSEs and attempt to correlate the findings. For instance, a network-based intrusion detection system might analyze the message activity from geographically related zones. Through correlation techniques, network-based intrusion detection might more accurately identify misused certificate as well as patterns of certificate misuse.

Because network-based intrusion detection is performed by back-end systems that can more easily scale than resource-limited RSEs, more detailed or computationally intensive intrusion detection analysis can be undertaken. In addition, while RSE intrusion detection capabilities need to be closer to real-time in operation, network-based intrusion detection has the benefit of processing data off-line.

For instance, network-based intrusion could perform correlation and analysis of certificate watch lists from multiple vehicles within an RSE zone. This technique takes advantage of the likely probability that there will be many observing vehicles whenever any vehicle generates a safety message. Each of the neighboring vehicles will have some knowledge about the local roadway environment and will perform their own evaluation about the legitimacy of a safety message. Each neighboring vehicle can report back to either the RSE or a network-based 1609.2 intrusion detection system with information about potentially misused certificates. The back-end system collects all the certificate watch lists and then performs off-line correlation and anomaly detection using statistical analysis.

This approach relies on the assumption that the majority of vehicles in the vehicle network are well intentioned. Under this assumption, responses from well-intentioned vehicles should outnumber those from misbehaving vehicles. Even if some of the responses from well-intentioned vehicles are wrong, sometimes because of inadequate data, the assumption is that the majority vote will be correct. All certificates may have

some ambient level of “believed” misuse. However, if one or more certificates significantly stand out, it is likely they are being misused.

Another way of using network-based intrusion detection is to perform off-line analysis of DSRC message captures. A message capture application could be installed on RSEs that would collect and store all broadcast messages and then upload them to a network-based system for post capture analysis. Message captures and more detailed analysis might be used as second tier efforts that get triggered in the RSE by the detection of suspicious behavior. Alternatively, network-based systems can operate continuously by randomly sampling RSEs or focusing on those where high rates of misused certificates have been historically or recently detected.

Network-based intrusion detection systems could also work in conjunction with the 1609.2 certificate authority to make use of information about re-keying totals and the association of certificates in assigned sets.

6.4 Intrusion Detection by Certificate Authority

With the exception of the re-keying process, the certificate authority is not in a forward position to observe vehicle communications and detect malicious messages or the misuse of certificates. It instead uses information about misused certificates to eliminate misbehaving vehicles from the vehicle network by publishing CRLs and denying re-keying requests.

The intrusion techniques used by the certificate authority primarily attempt to detect anomalies in certificate management data. The certificate authority maintains two important pieces of information for intrusion detection. First, the certificate authority maintains a count of the number of times a vehicle has replenished its certificates. Second, the certificate authority knows the set of certificates that have been assigned to each vehicle. Telcordia Technologies (Di Crescenzo et al. 2008) proposed a statistical model where misbehaving vehicles can be distinguished from the population by tracking the number of misused certificates for each vehicle in an anonymous manner, which takes advantage of the fact that the set of n certificates held by a vehicle is a relatively unique vehicle pseudonym or “anonymous ID.” A novelty of this approach is that the number of innocent vehicles that are wrongly accused can be controlled by dynamically adjusting the threshold. Analyses similar to this can be used with the concept of certificate watch lists that are reported by vehicles.

Knowledge of the set of certificates assigned to each vehicle can be maintained by the certificate authority. This information can be combined with knowledge of misused certificates to identify certificates that may be at high. If an intruder has successfully accessed one certificate on a vehicle, it is probable that other certificates on the vehicle have also been compromised. Therefore, the other certificates in all certificate sets that contain a misused certificate might be at higher risk than the remainder of the certificate pool. This information could be used to focus more the intensive intrusion detection capabilities at RSEs on a smaller set of certificates. This would conserve RSE resources or

permit the use of more resource-intensive techniques on a small set of certificates that could otherwise not be applied to all certificates.

The certificate authority does directly interact with vehicles during the re-keying process. The certificate authority has an opportunity at this time to detect intruder activities. One method that has been previously mentioned by Tengler et al. (2007b) involves the verification of a vehicle configuration hash. The certificate authority would keep record of a configuration hash that represents the unique vehicle identity. The OBE would need to generate this configuration hash from queried vehicle data and report it back to the certificate authority during the re-keying process. If the configuration hash does not match, it is assumed that tampering with the OBE or vehicle has occurred. In general, any authentication failure that occurs during the re-keying process could be a sign of intruder activity.

6.5 Intrusion Detection by Application Server

Vehicle network application servers can use many intrusion techniques that are commonly used for Internet-based applications. These include:

- Validating application input, commands, and command operands
- Establishing user activity behavior profiles
- Maintaining and analyzing detailed security logs and audit trails

In addition, many applications will require vehicles to identify and authenticate to back-end servers. Access controls for authentication and fine-grain authorization can be implemented for identifying services. Multiple simultaneous logins could be disallowed or tracked to identify the account involved. Finally, some applications may offer location-based features, which can be used to determine if user activity is occurring from outside expected geographic areas.

7 Conclusion

Securing vehicle networks is a complex challenge. Threat agents with different levels of resources and motives will seek to attack and exploit vehicle OBE, wireless communications, RSEs, and back-end network systems. Securing communication, preserving vehicle privacy, and maintaining the integrity of the system are three key objectives in a vehicle network security architecture. Safety applications require message authentication and integrity protection. PKI is recognized as the principal security technology needed for vehicle networks. Two basic certificate management approaches are shared certificates and unique certificates. The choice of certificate management scheme determines the balance of vehicle privacy and the ability of a vehicle network to remove bad actors. Efficient and effective intrusion detection is critical to maintaining system integrity. Intrusion detection capabilities will need to be deployed in all major components of a vehicle network.

References

- Di Crescenzo G, Pietrowicz S, Van Den Berg E, White R, Zhang T (2008) Vehicle segment certificate management using shared certificate schemes. US Patent Application 20080232583. www.uspto.org
- Di Crescenzo G, Zhang T, Pietrowicz S (2010) Anonymity notions and techniques for public-key infrastructures in vehicle networks. In: Wiley Inter Science Security and Communications Networks
<http://www.icann.org/tlds/org/questions-to-applicants-11.htm>
<http://www.intelldrivewayusa.org/library/research-reports/technical/dsrc-poc.php>
- IEEE P1609.2 (2006) Trial-use standard for wireless access in vehicular environments – security services for applications and management messages
- Pietrowicz S, Shim H, Di Crescenzo G, Zhang T (2008) VDTLS – providing secure communications in vehicular networks. INFOCOM 2008, Phoenix
- Pietrowicz S, Zhang T, Shim H (2010) Short-lived, unlinked certificates for privacy-preserving secure vehicular communications. In: ITS world congress, Busan
- Telcordia Technologies (2007) VII vehicle segment certificate management
- Tengler S, Andrews S, Heft R (2007a) Digital certificate pool. US Patent Application 20070223702. www.uspto.org
- Tengler S, Andrews S, Heft R (2007b) Security for anonymous vehicular broadcast messages. US Patent Application 20070222555, www.uspto.org
- van den Berg E, Zhang T, Pietrowicz S (2009) Blend-in: a privacy-enhancing certificate-selection method for vehicular communication
- White R, Pietrowicz S, Van den Berg E, Di Crescenzo G, Mok D, Ferrer R, Zhang T, Shim H (2009) Privacy and scalability analysis of vehicular combinatorial certificate schemes. In: 2009 IEEE CCNC, Las Vegas, 10–13 Jan 2009

49 Security, Privacy, Identifications

William Whyte

Security Innovation, Wilmington, MA, USA

1	<i>Overview</i>	1219
2	<i>Overview of Communications Security Services</i>	1221
2.1	Standard Cryptographic Security Services: Confidentiality, Authenticity, Integrity, Non-repudiation	1221
2.2	Anti-Replay and Relevance	1222
2.3	Availability	1223
2.4	Authorization	1224
2.4.1	Privilege	1224
2.5	Privacy	1225
2.5.1	Regulatory Aspects of Privacy	1226
2.5.2	Mechanisms	1227
3	<i>Standards, Field Operational Tests, Research Projects</i>	1227
3.1	Standards	1227
3.1.1	USA	1228
3.1.2	EU	1228
3.1.3	ISO	1230
3.2	Research Projects	1230
3.2.1	US	1230
3.2.2	EU	1231
4	<i>Security Services and Mechanisms for Applications and Their Support in Standards</i>	1231
4.1	Cooperative Awareness	1232
4.1.1	Mechanisms for End-User Privacy in Cooperative Awareness Systems	1234
4.1.2	Support in Standards	1236
4.2	Static Local Hazard Warning	1237
4.2.1	Support in Standards	1238
4.3	Interactive Local Hazard Warning	1238
4.3.1	Support in Standards	1240

- 4.4 Area Hazard Warning 1240
 - 4.4.1 Support in Standards 1242
- 4.5 Advertised Services 1243
 - 4.5.1 Support in Standards 1246
- 4.6 Local High-Speed Unicast Service 1247
 - 4.6.1 Support in Standards 1248
- 4.7 Local Groupcast Service 1248
 - 4.7.1 Support in Standards 1249
- 4.8 Low-Speed Unicast Service 1250
 - 4.8.1 Support in Standards 1251
- 4.9 Distributed (Networked) Service 1251
 - 4.9.1 Support in Standards 1252
- 4.10 Multiple Applications 1252
- 5 Security Management Services 1253**
 - 5.1 Initialization 1254
 - 5.1.1 Public Key 1255
 - 5.1.2 Symmetric Key 1255
 - 5.2 Key Update, Revocation, and Removal of Misbehaving Devices 1256
 - 5.3 Implementation and Support in Standards 1257
 - 5.3.1 Device-Side Case Study: VIIC Proof of Concept 1258
 - 5.4 Multiple CAs and Roaming 1261
 - 5.5 Infrastructure-Side Case Study: National ITS Architecture (USA) 1261
 - 5.6 Privacy Against the CA 1262
- 6 Multi-application, Physical, Platform, and IVN Security 1263**
 - 6.1 Vehicular Platforms 1263
 - 6.2 Non-vehicular, Mobile, Multi-application Platforms 1264
 - 6.3 Security for the Long Term 1265
- 7 Conclusion 1265**

Abstract: An ITS system will feature many different types of application. These applications all have their own security and performance needs, which may differ from each other. Additionally, the fact that different applications may coexist on the same device introduces additional security considerations. This chapter reviews the security mechanisms that may be used for different classes of application, and for the device as a whole, and surveys their deployment history and their support in standards. The aim is to provide an implementer of any ITS application with a usable starting point to help them determine which security services to use in their application and how those services should be implemented. Particular attention is paid to issues of privacy: ITS applications have an inherent risk of revealing personal data, such as current location, to parties who have no right to that data, and as such an implementer must take care to ensure that privacy is preserved to at least the level required by local regulations. The chapter also reviews security management operations such as issuing and revoking digital certificates.

1 Overview

The transportation system is vulnerable to many threats. An attacker who wants to disrupt traffic can throw caltrops from a bridge; an attacker who wants to distract the emergency services can make a hoax phone call; an attacker who wants to cause a specific vehicle to crash can shine a light in the driver's face; an attacker who wants to track a person can trail them physically, or put a tracking device on their car; an attacker who wants to tell whether a target has visited a certain spot (a liquor store, a motel to have an affair) can put a camera at that spot. As Intelligent Transportation Systems are introduced, the attacker's goals do not necessarily change, but his abilities do. A conventional attacker is limited in his ability to use the ITS system to cause damage to a single vehicle, but if all vehicles are networked, the attacker gains undetectability (not having to be physically present at the location of the attack) and scalability (being able to mount an attack in many places at once). Proper attention to communications security can make the difference between success and failure for the system as a whole.

The high-level goals of the system designers are:

- Provide assurance that messages received are accurate – they make true statements about the sender and about driving conditions.
- Ensure privacy for members of the public, partly to protect their rights, partly because if there is suspicion that the system is used to track people, then people will disable it and the benefits in saved lives, time, and money will be lost.
- While respecting privacy, ensure law enforcement access if necessary, and if proper procedures are followed.

This chapter focuses mainly on attacks mounted over the air (as an attacker who has physical access to a target vehicle can already mount a wide range of very serious attacks). An attacker may try directly attacking applications:

- Sending false application messages
- Blocking application messages

Or, since there will have to be some kind of security management system, the attacker may try to attack that instead:

- Sending false security management messages to ITS stations
- Sending false security management messages to the central security management entity
- Blocking valid security management messages

In discussing the design, this chapter focuses first on over-the-air attacks on applications and how they can be prevented, as this is the area of greatest interest to application developers; then on the over-the-air aspects of security management; then on the physical security of the devices themselves.

This chapter is structured as follows.

➤ [Section 2](#) provides a general overview of communications security services and outlines specific mechanisms that may be used to provide that service, as informative background to the rest of the discussion in this chapter. This includes discussion of both technical and regulatory aspects of privacy, which is an important security consideration for ITS.

➤ [Section 3](#) summarizes the state of existing standards for ITS communications security, describing the services they provide.

➤ [Section 4](#) is focused on security services consumed during the course of operations of ITS applications. The section is aimed at implementers of ITS applications and services. It provides a categorization of ITS applications into groups with similar characteristics. For each application group the section describes:

- The unique threats to use of that application
- The security services that must be provided to secure use of that application
- Recommended mechanisms to efficiently provide the required security services
- Case studies of the use of secure applications from that group within field operational tests
- Support for the recommended security mechanisms in standards

An implementer who reads this section should gain a good background on the specific communications security choices that are appropriate for their application, and should know where to look to find an interoperable specification of security mechanisms.

➤ [Section 5](#) discusses the security management services in the infrastructure that provide ITS stations with the keys and information necessary for them to engage in the secure transactions described in ➤ [Sect. 4](#). These services will in general be provided by a service provider who is different from the implementer of the security-consuming application itself. The services described in this section are not yet in widespread deployment and in some cases have not completed standardization, so this section is necessarily focused more on concepts than on the concrete details of implementation.

➤ [Section 6](#) describes approaches to platform security, which is the aspect of security design concerned with correct functioning of a device. This section considers threats to the operation of the system that originate on the ITS-S or within the In-Vehicle Network.

Note on terminology: This chapter uses “ITS Station” or “ITS-S” to refer to the communications and processing associated with a single ITS entity: vehicle, person with a personal device, roadside. Occasionally vehicular ITS-S and roadside ITS-S will be referred to as “On-Board Equipment” (“OBE”) or Roadside Equipment (“RSE”) respectively, if this is the common usage in the context under discussion. A functional element on the Internet that provides ITS services but does not communicate directly over the ITS spectrum is not an ITS-S but a “management entity” or similar. The term “application” is used somewhat vaguely to mean “processing that occurs between the network stack and the user interface.” Different “applications” do different “kinds of things”: A categorization of applications is provided in [Sect. 4](#). A single application is not necessarily a single file of executable binary code, though that may be a useful abstraction for the reader to bear in mind.

2 Overview of Communications Security Services

2.1 Standard Cryptographic Security Services: Confidentiality, Authenticity, Integrity, Non-repudiation

Almost all applications that communicate need the basic cryptographic security services:

- Confidentiality – protecting data from being read by unauthorized recipients
- Authenticity – giving assurance that received data comes from a known origin
- Integrity – protecting data from being altered by unauthorized recipients (achieved in practice by ensuring that such alteration can be detected)
- Non-repudiation – perhaps better called “third-party provability”: the ability to prove to someone other than the sender or receiver of a message that the message came from its sender

These services can be provided by public key cryptography or symmetric cryptography. The distinctions between the two are explained in many places, such as (Schneier 1996). For purposes of this handbook, the important distinctions are that public key cryptography has longer processing time and larger packet size overhead than symmetric cryptography, so symmetric cryptography should be used where possible, but that symmetric cryptography is suited mainly to communications with a central service provider and may not work in a peer-to-peer ad-hoc network setting.

All applications in this chapter need authenticity and integrity checking. Both of these are provided by digital signatures (which use public key cryptography) or message authentication codes (which use symmetric cryptography).

Confidentiality, which is provided by encryption, is needed to ensure that communications between two parties remain private. In the ITS area, some applications are broadcast and as such are intended to be heard by all recipients, including recipients with whom the sender does not already have a relationship with. These applications’ messages do not need confidentiality. Unicast or groupcast applications do.

Non-repudiation usually comes up in a context where a message may be challenged and the purported sender wants to demonstrate that they did or did not send it. Non-repudiation can therefore be used to protect a sender against a malicious claim that they sent a message. For example, if Alice's financial service provides non-repudiation, no other party can create a message in which Alice appears to ask for money to be transferred. Conversely, in principle, if Alice did in fact create a message she cannot in theory deny it (In practice, true non-repudiation is difficult to achieve in any system. Non-repudiation depends on the sender having a cryptographic private key that only they know, and since that key has to be stored somewhere, a sender under suspicion can always claim that the key was not stored securely enough and a forger got access to it. Depending on the context of the accusation and the properties of the platform where the key was stored, this has the potential to be a successful defense. Finally, non-repudiation has not been tested to any significant extent in a court of law and it is not clear where it stands legally).

In the ITS area, different applications will have different requirements for non-repudiation. Consider for example a car that is sending consistently inaccurate "heartbeat" messages. The car needs to be fixed, which is an inconvenience for the owner. Before putting the owner through that inconvenience, the authorities should be sure that the inaccurate messages actually came from that car and were not forged by a harasser. This proof is stronger if the messages have non-repudiation. As discussed later in this chapter, this is one of the arguments in favor of using public key cryptography to sign heartbeat messages, despite the performance and packet size overhead.

2.2 Anti-Replay and Relevance

If an attacker can record a valid message and, without altering it, replay it to the original or some other recipient, this can often be an attack. For example, if Bob records Alice sending her bank the message "pay Bob \$100," and if the bank does not check for duplicate messages, Bob can replay the message and get paid multiple times. To take a more contrived example, if a protocol has a message that means "toggle encryption on or off," then an attacker who can replay that message can change an encrypted session into an unencrypted session. In many cases, the data payload itself naturally includes checks that prevent replay (for example, in real life, a payment transfer message would typically include a timestamp). However, it is good practice for a cryptographic protocol designer not to assume that the data payload provides those checks; typically, the data payload is out of the control of the cryptographic protocol designer and future versions may lose protections that the system has been implicitly relying on for security. Note also that, depending on the setting, a replayed message may be intended as a confirmation of the previous copy of the message, rather than as an instruction to repeat the previous action. So the reaction to a replay will be highly dependent on the application.

Protection against replay may mean protecting literally against repeated messages. It may also mean protecting against taking action on a message that was generated

in the past but not received for some time. For example, “move 5 ft to the left” means one thing when organizing a group photo and another thing when standing just to the left of a cliff-edge. This type of attack is an attack on *relevance* – attempting to persuade a victim that an irrelevant message is in fact relevant now.

The two attacks are often considered together because they are addressed by similar mechanisms. Mechanisms to address these attacks include:

- **Timestamping:** Messages are integrity-checked and include a generation time. The recipient discards messages that are too old. Identical messages with different generation times are considered different messages. The anti-replay mechanism discards duplicate messages with the same timestamp. Care must be taken that the time counter is not going to roll over in the likely lifetime of the system. This method is most suitable for settings where there is not a continuous communication session, or the transmission medium is lossy; it allows messages to be received out of order, although if this is a desired property of the system, the application also must be designed to handle out-of-order messages.
- **Sequence numbers:** Messages in sequence are integrity-checked and include sequence numbers. The recipient only accepts a message if it has a higher sequence number than seen before. Care must be taken that the sequence number does not roll over in the likely lifetime of the system. This method is most suitable for settings where communications are infrequent but the connection is fairly reliable, and where timestamps are not suitable (for example, because the clocks on the communicating devices cannot be guaranteed to stay in synch).
- **Session commitment:** Messages in sequence are integrity-checked and include a commitment to the previous messages in the sequence, for example a cryptographic hash of the entire session to date. This is the approach used in Secure Sockets Layer (SSL)/Transport Layer Security (TLS) ([Wikipedia b](#)). This is most suited for an ongoing communications session over a non-lossy transport mechanism.

In an ITS system, specific location-aware applications will in general also need to be time-aware: An incident is something that happens at a particular time and place. This will naturally tend to give applications a level of protection against replay attacks, but the security design should not rely on this.

2.3 Availability

In addition to inserting false messages, an attacker may be able to attack a system by preventing messages from getting through. For example, consider an attacker who blocks the response to a message like “I’m going to transfer money from your account unless I hear otherwise in the next hour.” In practice, it is often not possible for the systems designer, especially for mass-market, commercial systems, to guarantee that a message will get delivered. The communications security designer must, however, contribute

to good design in this area, by ensuring that the system does not assume correct delivery and fails gracefully if messages are not successfully delivered.

Attacks on availability are commonly known as denial-of-service attacks. In general, denial-of-service attacks on wireless communications are hard to prevent, as a determined attacker can simply jam the signal at the appropriate frequencies. These physical attacks are mitigated by the fact that they are expensive to mount and relatively easy to detect. However, the system designer must take care to ensure that the system does not provide opportunities for force-multiplier denial of service attacks. In a force-multiplier denial of service attack, the attacker is able to use a small initial message to create a large amount of traffic. For example, consider a broadcast environment where an attacker is able to request an ACK to a message; a single from the attacker may attract hundreds of responses, and it is the response rather than the original message that overwhelms the system.

2.4 Authorization

Authorization is a security service that demonstrates that the entity taking an action is entitled to take that action. It builds on authentication (demonstrating that the sender is a valid unit) and integrity/replay protection (demonstrating that the sender sent “this message, here, now”). The demonstration of the sender’s permissions can be either *explicit* (also called *direct*) (so the permissions are directly encoded in an authorization token) or *implicit* (also called *indirect*) (so the authorization token contains some kind of identity, which the receiver can then map to a set of permissions). Examples of authorization include: demonstrating that the sender is a valid toll plaza; demonstrating that the sender is a specific tolling service customer; demonstrating that the sender is an emergency vehicle responding to an incident.

The authorization statement may be based on symmetric cryptography (in which case it is conventionally referred to as an authorization ticket) or on public key cryptography (in which case it is conventionally referred to as a digital certificate). Additionally, an authorization statement may say that the sender is allowed to take the given action under certain conditions. For example, a school bus may be converted into an evacuation vehicle on a particular day, or a police car may be authorized for hot pursuit within a particular county (The classic example here, of course, is Hazzard County). The authorization statement has “time constraints” or “geographic constraints.”

2.4.1 Privilege

Within an application, different senders may want to claim different classes of permissions, depending on the results they are intended to produce. The concept of privilege provides a useful framework for thinking about permissions that may be ranked by how

powerful they are. Privilege is a different concept from time-criticality; a message with high privileges may be of low time-criticality and vice versa. For example:

- V2V safety messages from end-user vehicles are providing information. The receiving unit may choose to alert its driver. The driver may choose to take action. An Emergency Brake Light message is highly time-critical but not particularly highly privileged.
- Emergency response vehicle V2V messages are higher privilege because there is a greater chance that the receiving unit will raise an alert to the driver, and that the driver will take some action in response to them. Most descriptions of the system distinguish between Private Mobile Users (PMU) and Public Service Mobile Users (PSMU) (US Department of Transportation 2011) to make this difference in roles as clear as possible.
- Emergency response traffic signal phase change requests are higher privilege still, because they result in action being taken automatically. (Note that they are not particularly time-critical – there is often a window of several seconds within which they can be received and still allow the signal to operate correctly).
- Within a tolling system, key update messages are higher privilege than simple tolling messages because they change fundamental system properties. But they are of lower time-criticality because a key update can usually be done at a time of the system's choosing.

High privilege messages need greater levels of assurance than lower privilege messages. Within a given application, there may be senders with different privilege levels, and the security system must provide a means to manage these privilege levels.

2.5 Privacy

A final property (not really a service as such) that systems can provide is privacy. Broadly speaking, “privacy” encompasses the concept that a person owns data relating to them, unauthorized parties should not be able to make use of personal data, authorized parties should only be able to make use of personal data with its owner's knowledge, and people should be able to choose which of their data they reveal to which authorized party (for example, their credit card company does not need to know their library card number, their library does not need to know their credit card number, and someone they pay cash to does not even need to know their name). However, the issue is complicated by considerations of traffic analysis and data mining: If someone can tell that two messages come from me, they can potentially learn information from the combination of the messages. This is often used for good purposes in legal investigations, where the pattern of a suspect's communication can reveal whether or not the suspect is part of a wider conspiracy. It is often a problem in, for example, healthcare research, where the combination of (medical condition, year of birth, area of residence) can often identify a single person, making it difficult to provide appropriately anonymized data sets to researchers.

Privacy also depends on the type of user. A person often has an expectation of privacy; a traffic signal does not. An ambulance requesting a signal change is requesting a high level

of privileges and so should have a low expectation of privacy, because the claimer of a high level of privileges should be accountable for their actions. On the other hand, the ambulance *driver* has an expectation of privacy.

Additionally, whatever privacy services are provided, there may be a legal right for appropriately authorized parties to reverse that privacy for law enforcement or to ensure the correct operation of the system.

2.5.1 Regulatory Aspects of Privacy

Different jurisdictions require different privacy properties of their systems. ITS systems offering cooperative awareness present a particular test case for privacy, as cooperative awareness depends on vehicles sharing information about their current position and other dynamic properties, and the set of cooperative awareness messages from a vehicle reveal the full path that vehicle took. A full set of cooperative awareness messages, combined with knowledge of the driver, therefore reveals the behavior of an individual. But cooperative awareness messages are the key to a lot of the safety-of-life benefits of the system. So a jurisdiction's privacy policies must balance the safety-of-life benefits with the privacy implications of cooperative awareness.

EU. The EU Data Protection Supervisor issued an Opinion relating to ITS (European Data Protection Supervisor 2010) which focuses more on services that are coordinated by a controller than on cooperative awareness. The key passages in this Opinion are:

- The use of anonymous data where appropriate . . . will not solve all data protection concerns as many data collected and exchanged through ITS may qualify as personal data. For the processing of personal data to be done on an anonymous basis, there must be no possibility for any person at any stage of the processing . . . to link the data with data relating to an identified individual, otherwise such data constitute personal data.
- The EDPS favors . . . that ITS services are provided on a voluntary basis. This entails that users must be able to freely consent to the use of the system and to the particular purposes for which it will be used. When the service provided relies on location data, appropriate information must be provided to the user . . . who must be in a position to withdraw this consent.
- The processing of location data should be strictly limited to persons acting under the authority of the provider of the public communications network or publicly available communication service or of the third party providing the value added service.

On the face of it, this Opinion makes the deployment of cooperative awareness systems very difficult, although there are ways to interpret it as permitting the transmission of cooperative awareness messages, for example by interpreting it as restrictions on processing by the receiver rather than the sender, or by allowing a user to opt out of sending at any time of their choosing for a limited time. At the time of writing, the exact status is not resolved.

USA. The USA does not have a Privacy or Data Protection officer with the same remit as the EU, so privacy policy is created in a more piecemeal fashion. The guiding law for driver privacy is the Driver Privacy Protection Act (18 U.S.C. § 2721 et. seq. 1997), although this is not ITS-specific. The closest thing to an official policy on privacy was issued by the National VII Coalition in 2007. It is not a policy statement but “a statement of intent from which specific rules and policies can be derived for the purposes of governing privacy protection, in the context of a National VII Program.” It “emphasizes collection and use of anonymous information that is not linkable to any individual *whenever possible*” (emphasis added). It differs from the EU in stating that vehicular data is in essence non-personal (“For example, in the future if a National VII Program enables vehicles to exchange anonymous positional and other vehicle information to help avoid collisions, the vehicles communicating are not personal information subjects, to the extent that the vehicles communicate only anonymous, impersonal vehicle data and do not communicate personal information associated with an identifiable individual.”) although this may be only a difference in emphasis.

In summary, as with the EU, at the time of writing, there is no definitive statement of the exact privacy protections that must be put in place for legal operation of cooperative awareness applications.

2.5.2 Mechanisms

Privacy as an over-the-air service is provided by a combination of different mechanisms. For example, any message containing personally identifying information must be encrypted for a recipient who is known to be authorized; conversely, any broadcast message must not contain personally identifying information.

This much is trivial, but for different types of attacker the mechanisms needed to provide privacy become more complex. Since an attacker may be able to record radio messages at different places, all identifiers within the message must change periodically. Since the attacker may record at the exact time that a change takes place, all the identifiers must change at the same time.

This chapter will discuss privacy mechanisms for each set of applications under that application, and privacy issues for multi-application systems in the final section.

3 Standards, Field Operational Tests, Research Projects

3.1 Standards

The following standards directly address communications security issues for ITS stations. This section lists and describes the existing standards. The next section describes how these standards may be used in practice for specific types of application.

3.1.1 USA

The USA's National ITS Architecture draws on standards developed in many different organizations. Architecture Development Team, 2007, identifies the security services needed by 22 ITS subsystems. Within the architecture, the main standards for application communications are defined by IEEE and the Society of Automotive Engineers (SAE). An architecture for infrastructure-side security management services is provided by the ITS "Core System," which at the time of writing is specified only to the Concept of Operations level.

IEEE

IEEE Standard 1609.2, *Wireless Access in Vehicular Environments (WAVE) – Security Services for Applications and Management Messages*, (IEEE Vehicular Technology Society 2006, 2011a), is a specification of size-optimized secured messages using public key cryptography based on elliptic curves. It has been used as the basis for many projects and field operational tests including SimTD (Deutschland 2009) and VIIC Proof of Concept (2008). At the time of writing it is completing a revision which will be published in late 2011 or early 2012.

IEEE Standard 1609.2 defines a template for *security profiles*; these are an interface that other standards organizations may use to define how their applications or services use 1609.2; 1609.2 itself contains a security profile for the Basic Safety Message from SAE J2735. SAE J2745 contains security profiles for other SAE-defined messages.

IEEE Standard IEEE 1609.11-2010, IEEE Standard for Wireless Access in Vehicular Environments (WAVE) – Over-the-Air Electronic Payment Data Exchange Protocol for Intelligent Transportation Systems (ITS), defines a protocol for electronic payment systems, including security. This is based on the framework given in ISO 14096 and ISO 15628.

SAE

SAE J2735 (2009) and J2945 (Society for Automotive Engineers n.d.) provide descriptions of how certain basic safety messages are to be encoded and secured.

3.1.2 EU

EU standards in ITS are primarily being formed through ETSI and CEN.

ETSI

ETSI documents have two identifying numbers, a work item reference and an ETSI document number.

ETSI documents come in a number of flavors, including TS (Technical Standard), ES (European Standard), and TR (Technical Report).

Document numbers	Title	Status (June 2011)
Doc. Nb. TS 102 731	Intelligent Transport Systems (ITS); Security; Security Services and Architecture; Security Services and Architecture	Published
Ref. DTS/ ITS-0050001		
Doc. Nb. TR 102 893	Intelligent Transport Systems (ITS); Security; Threat, Vulnerability and Risk Analysis (TVRA); ITS; Security; Threat Vulnerability and Risk Analysis	Published
Ref. DTR/ ITS-0050005		
Doc. Nb. TS 102 723-7	Intelligent Transport Systems; OSI cross-layer topics; Part 7: Interface between security entity and access layer; Cross-layer topics	Drafting stage
Ref. DTS/ ITS-0050007		
Doc. Nb. TS 102 723-8	Intelligent Transport Systems; OSI cross-layer topics; Part 8: Interface between security entity and network and transport layers; Cross-layer topics	Drafting stage
Ref. DTS/ ITS-0050008		
Doc. Nb. TS 102 723-9	Intelligent Transport Systems; OSI cross-layer topics; Part 9: Interface between security entity and facilities layer; Cross-layer topics	Drafting stage
Ref. DTS/ ITS-0050009		
Doc. Nb. ES 202 910	Intelligent Transport Systems (ITS); Security;; Identity Management and Identity Protection in ITS	Drafting stage
Ref. DES/ ITS-0050010		
Doc. Nb. TS 102 867	Intelligent Transport Systems (ITS); Security; Stage 3 mapping for IEEE 1609.2; IEEE 1609.2 profile	Drafting stage
Ref. DTS/ ITS-0050013		
Doc. Nb. TS 102 940	Intelligent Transport Systems (ITS); Security; Security architecture and ITS Station Security Management; Security architecture and Management	Drafting stage
Ref. DTS/ ITS-0050014		
Doc. Nb. TS 102 941	Intelligent Transport Systems (ITS); Security; Identity, Trust and Privacy Management; Identity, Trust and Privacy Management	Drafting stage
Ref. DTS/ ITS-0050015		
Doc. Nb. TS 102 942	Intelligent Transport Systems (ITS); Security; Access Control and Secure and Privacy-preserving services; Security; Access Control	Drafting stage
Ref. DTS/ ITS-0050016		

Document numbers	Title	Status (June 2011)
Doc. Nb. TS 102 943	Intelligent Transport Systems (ITS); Security; Confidentiality services	Drafting stage
Ref. DTS/ ITS-0050017		
Doc. Nb. TR 102 893	Intelligent Transport Systems (ITS); Security; Threat, Vulnerability and Risk Analysis (TVRA); TVRA Revision	Drafting stage
Ref. RTR/ ITS-0050018		

3.1.3 ISO

ISO TC 204 WG 16 is focused on Continuous Access for Land Mobiles (CALM), meaning continuous communications for vehicles. At the time of writing its security activities included the following. More details are available at www.calm.hu.

Item	Title	Status
11766	Intelligent transport systems – Communications access for land mobiles (CALM) – Security considerations for lawful interception	Issued 2010
11769	Intelligent transport systems – Communications access for land mobiles (CALM) – Data retention for law enforcement	Issued 2010
13181-1	CALM Security Part1: Framework	In process
13181-2	CALM Security Part2: Threat Vulnerability and Risk Analysis	In process
13181-3	CALM Security Part3: Objectives and Requirements	In process
13181-4	CALM Security Part4: Countermeasures	In process

3.2 Research Projects

3.2.1 US

There have been a number of large-scale, government-funded research and field test projects that have included security, most notably VSCC (2003–2006) (Vehicle Safety Communications Consortium 2005), VSC(A) (2006–2009) (Vehicle Safety Communications Applications n.d.), VSC3 (2010-present), and the VIIC Proof of Concept (2008). The forthcoming Safety Pilot project (scheduled to start in late 2011) will also include many security-aware ITS applications. Where relevant, the conclusions of completed projects are noted in the analysis below.

3.2.2 EU

eSafety is a European Commission-initiated body tasked with determining how to accelerate the deployment of ITS systems for safety of life. It sponsors technical research projects along with projects on policy and has a working group specializing in security.

A number of EU research projects have been funded under the Sixth and Seventh framework specifically into ITS security: Sevecom (2006; Kung et al. 2007) for general ITS communications security; Preciosa (Preciosa Project n.d.) for considering privacy; Evita (Evita Project n.d.) for secure in-vehicle communications; Oversee (Oversee Project n.d.) to address platform security; and Preserve (Preserve Project n.d.), to take output from the previous projects and integrate and field test their results to bring them to a pre-deployment state.

There have also been a number of field tests specific to individual countries, such as SimTD (Deutschland 2009) in Germany.

Where relevant, the conclusions of completed projects are noted in the analysis below.

4 Security Services and Mechanisms for Applications and Their Support in Standards

The appropriate security services for an application depend on the application's communications patterns: what types of entities send and receive application messages. For each application, the security services that it needs are a result of the threat model analysis. For a given security service, there may often be multiple different mechanisms that are capable of implementing it (for example, authentication can be implemented using public key or symmetric key cryptography; within each of public key and symmetric key cryptography, there is a choice of possible different algorithms). If different mechanisms may be used to implement a service, the choice of recommended mechanism will be driven by other factors such as performance, standardization, easy availability, and so on. Once an individual mechanism is identified as the best way to implement a service, the mechanism will be standardized. This distinction between services and mechanisms mirrors the distinction in ETSI between Stage 2 documents, which identify security services and are somewhat abstract, and Stage 3 documents, which specify particular mechanisms and are concrete and implementable.

This section is directly aimed at implementers who want to build platforms that support specific applications. For each application, this section uses the following structure:

- Unique features of the threat model
- Security services needed to counter the threats
- Mechanisms to implement those services that meet the security requirements
- Standards, if any exist, that specify those mechanisms

Security services for advertisements are discussed separately from security services for the services they advertise. The aim of the section is to point readers in the right direction so that they can deploy systems that are secure, efficient, and interoperable, and understand why those systems are specified the way they are.

Different organizations identify and group ITS applications in different ways: (Architecture Development Team 2007) identifies 22 “subsystems” and 16 “architecture flows”; (RITA Joint Programs Office 2011) identifies six categories of “interfaces”; SeVeCom (2006; Kung et al. 2007) identified 56 different applications in ten different clusters. The grouping here is based on ongoing work within ETSI and categorizes applications under the following headings. The headings overlap somewhat for ease of analysis.

- Traffic pattern: Unicast, broadcast, geocast
- Network mode: Multi-hop, single-hop
- Time-criticality: Critical, high, low
- Data size: small (single message), medium (multiple messages but transaction can be completed in the time it takes two vehicles to pass at high speed); large (larger than medium)
- Transmission frequency: Frequent (multiple times a second) or infrequent (once a second or less)
- Endpoints: V2V; V2I/I2V; Vehicle to remote infrastructure, reached over a backhaul network
- Session: Individual messages; unicast local session; unicast session with a server remote over the network; unicast session with a server remote over the network, which must be maintained across several V2I communications sessions

4.1 Cooperative Awareness

Cooperative awareness messages are: broadcast, single-hop, high time-criticality, small data, frequent transmission, V2V, no session. Based on single broadcast message with no explicit coordination. Examples: Emergency vehicle, slow vehicle, emergency brake lights.

Cooperative awareness applications are the most widely studied and in some ways the most challenging of all applications.

User types: Public service vehicle; end-user vehicle

Threat model notes:

- *Likelihood of attack:* Since cooperative awareness messages are issued by the most common device in the system, the OBE, there is a high likelihood that some device somewhere will be compromised and false messages will be introduced into the system.
- *Sybil attacks:* An attacker who compromises a unit may be able to generate multiple false messages, either pretending to be multiple vehicles in one place, or

transmitting as a single vehicle but in multiple places simultaneously, or both. The vehicle's authorization might have geographic constraints, but even so, since a vehicle will want to be authorized to send cooperative awareness messages over a wide area, the authorization constraints by themselves will not be enough to prevent the multiple-places attack. This means that a misbehaving unit, until it is detected, can have a much bigger impact on the system than a correctly behaving unit.

- *Risk due to false messages:* A false message is a risk in as much as it might cause a false alert to be raised to a driver. If a false alert is raised to a driver, he or she might take evasive action, which might in turn have a chance of causing an accident (though a less serious one than the one the driver thinks she's avoiding). In general, though, the risk of false messages in the system is not so much that they might immediately increase accidents, as that they might cause the system performance to degrade to a point where the system is no longer useful. This might come about simply by overwhelming the channel, or by saturating the driver with so many false alerts that they stop paying attention to warnings the system gives them.
- *Availability:* Availability is a particular concern for cooperative awareness applications. As the penetration rate of cooperative awareness increases, drivers may come to depend increasingly on the warnings from the cooperative awareness and to pay less attention to what they see on the road in front of them (a similar affect has been observed where the introduction of seatbelt laws appears to make drivers more reckless (Adams 1982), though this result is somewhat contested (Cohen and Einav 2003)). If an attacker can impact the availability of cooperative awareness messages in a world where drivers have come to rely on cooperative awareness alerts, the accident rate will go up; it may even go up relative to a world with no cooperative awareness messages at all.
- *Privacy:* ITS users are living in a world where there are already many threats to privacy. A person can be tracked by their cellphone radio transmissions. Cars have license plates. Security cameras are cheap and widespread. A determined attacker can attach a tracking device to a target's car, or even hire people to follow a target around. Nevertheless, the use of cooperative awareness systems does increase the risk that personal data will be exposed. In particular, an eavesdropper can invest in infrastructure to receive and record messages, enabling them to capture and data-mine telematics data for personal information. Although the cost of the first privacy attack for this attacker will be high, the marginal cost will be extremely low, and so this has to be considered a meaningful risk to privacy even in the presence of the other threats to privacy discussed above.

Services:

- *Confidentiality:* No need for confidentiality – messages are broadcast.
- *Authenticity:* Messages must be authenticated to prevent injection of false messages into the system. For V2V systems, public key cryptography with digital certificates

is the natural way of providing authenticity because this is the easiest way for a recipient to trust a message from a previously unknown sender.

- *Integrity*: Necessary, provided by digital signatures.
- *Authorization and privilege classes*: Emergency vehicles have different privileges from end-user vehicles. The authorization statement (digital certificate) may also want to specify the quality of the security implementation (how well protected the sender's keys are) and the expected accuracy of the inputs to the cooperative awareness message.
- *Non-repudiation*: Necessary for misbehavior detection so that the sender of a misbehaving message may be uniquely identified.
- *Anti-replay and relevance*: Relevance checking is necessary to prevent false warnings from being raised to drivers. This may be provided at the application layer or by the security services. Replay is confirmation.
- *Availability*: A cooperative awareness application should be aware of whether there is likely to be a significant threat to availability, be it malicious or accidental, and should ensure that the driver is alerted to pay more attention to the road.
- *Privacy*: When an emergency service vehicle is performing an emergency response task or otherwise claiming use of emergency service vehicle privileges, the vehicle should be identified. For non-privileged users, the issue of privacy is more complex and is addressed in the next subsection.

Performance and Mechanisms:

- Authenticity, integrity, authorization, non-repudiation are provided with public key cryptography, digital signatures, and digital certificates. The greatest performance concern is congestion on the channel. To alleviate this, most implementations use the Elliptic Curve Digital Signature Algorithm (ECDSA) (Johnson and Menezes n.d.) for cryptography. Public key cryptographic operations are slow: On a 400 MHz processor, ECDSA verifications at appropriate security levels can take up to 50 ms. In a likely real-world situation, a device may have to receive over 200 messages a second, and will have to determine whether or not to alert the driver in less than 100 ms from receipt of the message. There are two basic ways of handling the performance requirements: either using hardware acceleration, or only verifying certain messages.

4.1.1 Mechanisms for End-User Privacy in Cooperative Awareness Systems

The following techniques are commonly recommended to improve these systems' privacy:

- No identifiers in the message that directly reference a known real-world identity for the vehicle or driver.

- Identifiers change periodically (there is not yet a research consensus on how often identifiers should change).
- Whenever a single identifier in the message changes, all the identifiers in the message must change. This requires synchronization up and down the network stack to ensure that application level IDs, digital certificates, and source addresses at the MAC and (possibly) IP layer are all changed in a single interval between one message and the next.

A lot of this privacy-preserving behavior is implemented by local processing and has no interoperability implications. However, it must nevertheless be standardized in case different behaviors can be used to distinguish different vehicles.

The attacker discussed above, who can record data over a wide area, will be able to reconstruct paths even if identifiers change by simply “joining the dots” of the (location, velocity) provided in the cooperative awareness messages. To address this, researchers have explored various approaches:

- Users may choose to opt out of sending cooperative awareness messages at a time and for a period under their control.
- Vehicles may explicitly coordinate identifier changes with other vehicles (Gerlach and Guttler 2007).
- Vehicles choose to change identifiers during “silent periods” that are chosen so that with high probability, other vehicles will be in a silent period at the same time (Buttyán et al. 2009).

None of these approaches have yet been implemented in practice. Their likely impact is hard to judge. The benefits of cooperative awareness go, roughly speaking, as $p_s p_r$, where p_s is the fraction of vehicles that are sending messages within the system and p_r is the fraction of vehicles that are receiving messages. Without any silent periods or opt-out, $p_s = p_r$, but some group of drivers at the margin may choose to disconnect their cooperative awareness modules altogether, lowering p_r . With silent periods or opt-out, p_s will typically be less than p_r by a certain amount, but fewer users will disconnect their devices altogether. The optimal mix has not yet been determined.

Additionally, the legality of cooperative awareness is not entirely clear under the privacy regulations discussed above. Final dispositions have not been made at the time of writing, but it seems highly likely that, for deployment to be approved, regulators will require the ability for users to opt out of sending at any time of their own choosing.

Finally, note that to support law enforcement access, it may be necessary for certificates to include a masked identifier such that eavesdroppers cannot identify the vehicle but it can be identified for law enforcement or other enforcement purposes. Other factors affecting the design of anonymous credentials include the ability for vehicles to obtain new credentials, and the ability of the security management authorities to remove vehicles from the system using revocation lists.

4.1.2 Support in Standards

General:

- *Privilege classes*: Not specified in any standard.
- *Confidentiality*: Not needed.
- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.
- *Privacy*: No standardized solution. Different approaches have been tried in different field trials.

US:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: IEEE 1609.2 provides definitions of digital certificates and signed messages to give these services. In the USA, cooperative awareness is implemented with the Basic Safety Message (BSM) specified in SAE J2735. SAE has approved a *security profile* for 1609.2 specifying exactly how 1609.2 services are to be used for cooperative awareness. Note that the BSM timestamp will roll over in the lifetime of the system, so to prevent replay attacks, an additional timestamp is added by the security services.
- *Privacy: Certificate format*. A proposal for a standard for privacy-preserving certificates is anticipated by the end of 2011. *Unlinkability*: 1609.4 allows for the MAC address to be changed to provide unlinkability.
- *Performance*: SAE recommends that messages are only verified when necessary, i.e., when they will result in an alert being raised to the driver.

EU:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: Provided by the ETSI profile of IEEE 1609.2 in ETSI TS 102 867 (2011). The ETSI Cooperative Awareness Message (CAM) uses a timestamp which is unlikely to roll over in the lifetime of the system, and so there is no need to add an additional timestamp within the security services.
- *Privilege classes*: Not currently specified in any standard but under consideration within ETSI ITS.
- *Confidentiality*: Not needed.
- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.
- *Privacy*: No standardized solution yet.
- *Performance*: ETSI aggregates all CAM messages into a Local Dynamic Map, which may be used by multiple human interaction applications. As such it is hard to tell a priori which messages must be verified and which may safely go not verified. ETSI is recommending that all messages are verified and that the necessary performance is achieved by using hardware cryptographic acceleration.

4.2 Static Local Hazard Warning

Static local hazard warning messages are: broadcast, single-hop, high time-criticality, small data, frequent transmission, I2V, no session. Based on single broadcast message with no explicit coordination. Examples: Curve rollover, signal phase and timing, signal violation (when implemented by on-board logic reacting to signal phase and timing message), wrong way warning (when implemented by on-board logic reacting to information about legal traffic directions).

User types: RSEs send, OBEs receive.

Threat Model Notes:

- *Restricting attacks by authorization:* RSEs are often bound to a particular site by the authorization statement, in which case an attacker who compromises an attack can only send messages relevant to that site. If the RSE has frequent access to an authorization authority, it can also be issued with short-lived authorization statements. Both of these act to reduce the damage that can be done by an attacker who compromises a single RSE.
- *Risk due to false messages:* The risk due to false messages may be greater than for vehicle-based cooperative awareness due to user interface choices in the receiving vehicle. In the case of cooperative awareness, the only information a driver gets is an alert or a lack of alert. In the case of, for example, curve rollover warnings, a driver may be actively encouraged to drive faster than is safe if a false message misstates the curve parameters. In the case of traffic signal phase, a false message may encourage a driver to drive faster than they should to make it through the current green light, while in fact, the attacker has calculated the message so the light will be red when the driver reaches it and the driver will run it. So in contrast to cooperative awareness, where false messages are less effective than denial of service at causing accidents, with static local hazard warning, false messages may be more effective than denial of service.
- *Availability:* If roadside hazard messages are primarily presented to the driver as alerts, then availability may be an issue for the reasons discussed under cooperative awareness, i.e., that the driver comes to rely on hazard messages and drives more recklessly in their absence. However, note that not every hazardous location will be instrumented, so the system is more likely for this case than for vehicle-originating messages to be set up to tell the driver whether or not an alerting device is present in the first place. This reduces the risks due to diminished availability – drivers will be driving more alertly in the first place – but increases the risks due to false messages as discussed above.

Services:

- *Confidentiality:* No need for confidentiality – messages are broadcast.
- *Authenticity:* Messages must be authenticated. Public key cryptography with digital certificates is appropriate.

- *Integrity*: Necessary, provided by digital signatures.
- *Authorization and privilege classes*: Typically, there is one natural privilege class: RSEs sending local hazard warning messages. The authorization statement (digital certificate) may also want to specify the quality of the security implementation (how well protected the sender's keys are).
- *Non-repudiation*: Necessary to prevent receivers (vehicles) from masquerading as senders (RSEs).
- *Anti-replay and relevance*: Relevance checking is necessary. This may be provided at the application layer or by the security services. Replay is confirmation.
- *Availability*: It is helpful for the receive-side system to be aware of ongoing threats to availability.
- *Privacy*: RSEs have no requirement for privacy.

Performance and Mechanisms:

- As with cooperative awareness, the mechanism best suited to provide authenticity, integrity, authorization, non-repudiation is public key cryptography, digital signatures, and digital certificates using ECDSA. As with cooperative awareness, concerns about performance (message processing latency and throughput on the receive side) can be addressed either by using hardware acceleration, or by only verifying certain messages.

4.2.1 Support in Standards

General:

- *Confidentiality, Privacy*: Not needed.
- *Privilege classes*: Not specified in any standard. Privilege classes may not be necessary for these applications.
- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.

US:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: SAE J2945 ([Society for Automotive Engineers n.d.](#)) provides security profiles for 1609.2 for use with static local hazard warning applications.

EU:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: Not yet specified

4.3 Interactive Local Hazard Warning

Interactive local hazard warning messages are: broadcast followed by unicast, single-hop, high time-criticality, small data, infrequent (only when hazard exists), V2V or V2I/I2V, local session. Examples: Precrash, Intersection collision warning (when coordinated

by an intersection), cooperative glare reduction, signal violation (when implemented by signal logic reacting to cooperative awareness message from vehicle), wrong way warning (when implemented by roadside logic reacting to cooperative awareness message from vehicle).

The basic model for these applications is that station A receives a cooperative awareness message from station B and then returns a message to station B requesting that it takes a particular action. Based on this there may be additional data exchanges. These exchanges may contain more personal information than is included in cooperative awareness messages.

Threat Model Notes:

- *Privacy:* The threat to privacy from these applications depends on deployment patterns that have not yet been established. It seems likely that individual messages from these applications are greater threats to privacy than cooperative awareness messages because they require the vehicle to produce more unique information than CAM does, but that the privacy risk is relatively low because these reflect relatively rare events. An active attacker may be able to trick a victim into compromising privacy by initiating an interactive local hazard warning exchange.
- *Availability:* These are applications that will be used rarely so as well as issues with availability, when a driver is more careless in the absence of an alert, there are also issues with distraction, as a driver is confused about how to cope with an incoming message that they are not familiar with. The security implications differ between lack of availability because the service is not being provided, and lack of availability because the service is being swamped. The driver should be made aware of problems with the communications medium.

Services:

- *Confidentiality:* So long as these exchanges do not contain any more personal information than is included in cooperative awareness messages, confidentiality is not necessary.
- *Authenticity:* Messages must be authenticated. The initial authentication must be done using digital certificates. It is possible that extended exchanges can be protected with a negotiated symmetric key, to improve efficiency.
- *Integrity:* Required.
- *Authorization and privilege classes:* RSEs and OBEs will have different privilege classes. It is possible that there will be public safety OBEs. It may also be useful to distinguish between senders on the basis of the quality of their security implementation, or on the basis of the algorithm they use for (for example) cooperative glare reduction. Maybe both parties will need special privileges to run these applications in the first place.
- *Non-repudiation:* If these messages will be used for misbehavior reporting, they need non-repudiation. Otherwise not.
- *Anti-replay and relevance:* Messages must be time- and location-stamped. Replay is confirmation.

- *Availability*: Implementations may choose to make the driver aware of when availability is limited.
- *Privacy*: Messages must not contain any more information than is necessary to execute the application. Encryption may help preserve privacy but the gains will be marginal; more bang for the buck in removing personal information in the first place. Unlinkability will come from the same mechanisms as with CAM: all fields that are just identifiers should change from one use of the application to the next; all static fields that have meaning should be sanitized so they refer to a group of vehicles.

Performance and Mechanisms:

- For initial messages: authenticity, integrity, authorization, non-repudiation from public key cryptography, digital signatures, and digital certificates using ECDSA. Concerns about performance (message processing latency and throughput on the receive side) can be addressed either by using hardware acceleration, or by only verifying certain messages, or by using symmetric cryptography for ongoing exchanges.

4.3.1 Support in Standards

General:

- *Confidentiality*: Not needed.
- *Privacy: Certificate format*: Not yet defined, see cooperative awareness. *Unlinkability*: can make use of changing MAC addresses in 1609.4.
- *Privilege classes*: Not specified in any standard. Privilege classes may not be necessary for these applications.
- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.

US:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: SAE J2745 provides security profiles for 1609.2 for use with dynamic local hazard warning applications.

EU:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: Not yet specified

4.4 Area Hazard Warning

Area hazard warning messages are: broadcast, multi-hop with geocasting, high time-criticality, small data, infrequent (only when hazard exists), V2V or V2I/I2V, no session. Examples: Traffic condition warning (stopped vehicles more than one hop away).

This category is known as Decentralized Environmental Notification (DEN) within ETSI.

Threat model notes:

- *Security of geocasting:* This application category uses geocasting. Geocasting uses a source location and destination area, and the routing then consists of a series of individual forwarding decisions made by intermediate nodes based on the routing algorithm and their understanding of local conditions. Consider two different attacks: attacks on the source or destination location, and attacks on the intermediate routing decisions.
- *Attacks on source/destination:* If the destination area is not authenticated, the geonetworking protocol may be a means for an attacker to mount a force-multiplied denial of service attack by introducing a single message with a large or distant destination area that propagates uncontrollably throughout the network. This DoS/QoS attack must be countered not just by cryptographic mechanisms but by sensible routing protocol design that addressed the inherent congestion issues arising from forwarded messages.
- *Attacks on routing:* If an attacker can interfere with routing decisions, they can cause a local force-multiplier denial of service attack on all applications by increasing channel congestion, or a denial-of-service attack on only the current geocast message by redirecting it out of the system. However, to mount this attack, the attacker must be local and active, and if this is the case, they have a range of other options for mounting a denial of service attack. Adding authentication to the single-hop routing information increases the ability to create an audit trail for packets (assuming that the information can be recovered from the intermediate nodes) but does not appear to add significant real-time protection.
- *Privacy due to geocasting:* The geocasting routing protocol uses location information from local vehicles to make routing decisions. For geocasting to work properly, vehicles must be transmitting their location and velocity. Since this already is part of the cooperative awareness functionality, geocasting does not pose any additional security risk to a system that uses cooperative awareness.
- *Privacy due to messages:* Messages do not contain any more information than is contained in cooperative awareness messages.
- *Risk due to false messages:* DEN messages are less important than cooperative awareness messages; as with cooperative awareness, the risk that a false message will cause dangerous driving is extremely low.
- *Availability:* As the distance from the source increases, the risk also increases that a geocast message will not reach its destination successfully. Applications must be built with this in mind. As such, it is highly unlikely that drivers will come to rely on (multi-hop) DEN to the extent that they rely on (single-hop) cooperative awareness. Specific security services to guarantee availability for DEN are therefore not important.

Services:

- *Confidentiality:* No need for confidentiality.
- *Authenticity:* Messages must be authenticated. Public key cryptography with digital signatures and certificates is appropriate.

- *Integrity*: Messages must be integrity-checked. Again, public key cryptography with digital signatures and certificates is the appropriate mechanism.
- *Authorization and privilege classes*: DEN messages allow the sender to recommend a particular evasive action for receivers to take, and to state the severity of the event being reported. It may be desirable to set privilege classes to restrict the type of messages that more limited devices can send.
- *Non-repudiation*: Necessary for misbehavior reporting.
- *Anti-replay and relevance*: Relevance checking is necessary to prevent false warnings from being raised to drivers. This may be provided at the application layer or by the security services. Replay – in other words, a recipient receiving the same message twice – is inherent in the routing protocol and is not an attack.
- *Availability*: Availability over multiple hops from the origin is not a significant security risk, though as noted above protocol designers need to balance availability and channel congestion. Over single hops, this protocol has the same effect as CAM and is subject to the same availability considerations.
- *Privacy*: The originator of a message may be giving away personal information that would not be included in a CAM. If this is the case, sending DEN needs to be an opt-in process. Forwarding nodes do not give away any information other than the information included in a CAM.

Performance and Mechanisms:

- Authenticity, integrity, authorization, non-repudiation from the sender: Public key cryptography, digital signatures, and digital certificates using ECDSA. Concerns about performance (message processing latency and throughput on the receive side) can be addressed either by using hardware acceleration, or by only verifying certain messages.
- Authenticity, integrity, authorization, non-repudiation at the networking layer: As noted above, there is no real need for security services to be applied to the forwarding information. However, if desired, it may be included in the protocol and if this is the case the appropriate mechanism is public key cryptography, digital signatures, and digital certificates using ECDSA.

4.4.1 Support in Standards

US: No US standards support multi-hop.

EU:

- *Authenticity, integrity, authorization, non-repudiation, anti-replay, relevance*: Provided by the ETSI profile of IEEE 1609.2 in ETSI TS 102 867. There is no need to add an additional timestamp within the security services.
- *Privilege classes*: Not currently specified in any standard but under consideration within ETSI ITS.
- *Confidentiality*: Not needed.

- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.
- *Privacy*: No standardized solution yet.
- *Performance*: ETSI aggregates all DEN messages into a Local Dynamic Map, which may be used by multiple human interaction applications. As such it is hard to tell a priori which messages must be verified and which may safely go not verified, except that exact duplicates of messages already received may be discarded. ETSI is recommending that all messages are verified and that the necessary performance is achieved by using hardware cryptographic acceleration.

4.5 Advertised Services

Advertised services refer to services where a Provider unit sends out a message of a particular type advertising that the service is being provided, and a User unit with the corresponding user application connects to the service. This description is based on WAVE Service Announcements (WSAs) as described in IEEE 1609.3 (2010a).

Advertisements are not application messages themselves, though they may contain information allowing the user application to decide whether to connect. For example, a service advertisement for tolling might contain an identifier for the tolling operator.

They are broadcast and cannot be encrypted. They are typically sent multiple times a second.

In a lot of cases, a User unit will be an end-user vehicle with a strong expectation of privacy.

Example: tolling, local information download

Threat model notes:

- *What is being protected?* An advertisement is the start of a hand-off from one part of the communications process to another. It may be useful to demonstrate that the provider device is allowed to send the advertisement. However, this demonstration only protects the advertisement, not the service itself. If the service is protected and properly authenticated, then the main purpose served by protection on the advertisement is to prevent a User from inadvertently revealing information about themselves by responding to an advertisement.
- *Authenticity*: If an advertisement is authenticated, it means two things: (a) the Provider is authorized to run or enable access to the services in the advertisement; (b) the Provider is authorized to advertise those services. Secure access to the services may be provided by mechanisms within the service protocol rather than within the advertisement, so security property (a) does not need to be authenticated within the advertisement itself. However, property (b) can only be provided by authenticating the advertisement.

- *Privacy against eavesdroppers:* An eavesdropper in this context does not send advertisements but is able to overhear and record all advertisements and responses to them.
- *Application data:* Unless the response to an advertisement is encrypted, it may reveal identifying information. To combat this, service design should ensure that no identifying information is revealed until an encryption key is established.
- *The fact that a service is installed as an identifier:* If a service is relatively rare among Users, the fact that a User responds to its presence in an advertisement may be an identifying activity in itself. (As an example of a rare service, consider a toll road with multiple different operators such that each vehicle only responds to the operator it is registered with. If the different operators are advertised in different advertisements, a vehicle gives away its operator by which one it chooses to respond to. Or a service that foreign-registered cars respond to but domestically registered cars do not). To mitigate this, all identifiers should be changed between different responses to a service, although note that this is not a full protection against the attack; if the service is sufficiently rare, its presence on its own is an identifying activity. The only defense against an eavesdropper in this case is if the service is offered by few Provider units, so that the eavesdropper has few opportunities to observe a User response to that service. Another protection at the procedural level is to ensure that these services are opt-in (note that if they are rare, they are likely to be opt-in as well, although one can think of exceptions such as services that are restricted to high-end vehicles from a particular OEM); at that point, one can argue that the privacy threat is equivalent to the privacy threat from other opt-in services such as carrying a mobile phone.
- *The fact that multiple services are installed as an identifier:* A single service on the User may be too common to count as an identifier, but the use of multiple services may be a distinguisher (for example, if there are three services, each of which is on 10% of all vehicles (chosen independently), then the knowledge that a vehicle runs all three services pins it down to 1% of the total vehicle population).
- *Security against an active attacker (pretending to be a Provider):* An active attacker can actively generate advertisements. Assume that the attacker cannot actually provide a service unless they are authorized to do so. The attacker's ability is therefore simply to provoke responses to advertisements for services that are not in fact being offered. This gives the attacker three options:
 - *Poorly designed handoff:* If the service/advertisement handoff is badly defined, the attacker could obtain application data.
 - *Attack privacy with uncommon service:* The attacker can advertise rare services, allowing him to break privacy as discussed above.
 - *Force-multiplier attack on availability with common service:* If the attacker advertises a common service, such that almost all units respond to it, this may overwhelm the channel with little effort on the attacker's part.

- *No authentication on advertisements*: If a service is extremely widespread (so responding to it is not a privacy concern) and contains its own security at the application level, the Provider unit may not need to demonstrate that it is authorized to offer it.
- *Advertisements with multiple services*: An advertisement message may contain advertisements for multiple individual services. It could be that some of those services do not need to be authenticated (because they are very common and have application level security, as discussed above), and some do.
- *V2V Advertisements*: Advertisements may potentially be used to kick off precrash information exchange. If precrash information includes more identifying information than standard cooperative awareness messages, as seems likely, then an active attacker can possibly use false precrash warnings to harvest information from nearby vehicles. This attacker will have to have compromised an OBE, but as noted above, there will almost certainly be a number of compromised OBEs in the system at any one time. The defense against this attack is twofold: (a) plausibility checking at the application level to ensure that the Provider unit moves first in revealing its information and to ensure that the User unit only responds to precrash warnings that are plausibly close to it; (b) an enforcement system such that precrash warnings are automatically reported to the authorities, so that a false advertiser can be identified and removed as quickly as possible.
- *Provider Privacy*: If a Provider that is offering a service to end-user vehicles is an RSE, it has no expectation of privacy. If an OBE, it may have an expectation of privacy. Because (a) a compromised OBE can use service advertisements to compromise the privacy of other OBEs, and (b) this will in general not be geographically limited so a single compromised OBE can threaten the privacy of multiple OBEs across the system, an OBE advertising a service should have a lower expectation of privacy itself. At the very least, although it may not have an expectation that it will be trackable by an eavesdropper, it should expect that its advertisements will be collected by a central authority and occasionally analyzed in a privacy-compromising way. Also note that if the set of services advertised is persistent, that set becomes an identifier for the Provider, and even if the advertisement does not contain personal location data, if it is sent from the same MAC address as that vehicle's CAMs, it will allow tracking. Offering a Provider service should always be an opt-in activity (or at the very least something that an OBE can opt out of).

Services:

- *Confidentiality*: Not needed.
- *Authenticity*: If the advertisement is for a common service or combination of services, and if that service contains authentication at the application level, the advertisement itself does not need to be authenticated. If this is not the case, the advertisement does need to be authenticated. Since advertisements are broadcast, if they are authenticated, it needs to be using public key cryptography and digital certificates.

- *Integrity*: The advertisement needs integrity checking under the same circumstances as it needs authorization and authentication.
- *Authorization and privilege classes*: In the general case, a Provider should show authorization for being able to send advertisements. If the advertisement is for a common service or combination of services, the advertisement itself does not need to be explicitly authorized.
- *Non-repudiation*: If services need to be authenticated, they also need non-repudiation to protect valid Providers against malicious misbehavior reports.
- *Anti-replay and relevance*: Advertisements need relevance checks. Replay is confirmation.
- *Availability*: The attack on availability by an active attacker, above, can be mitigated by authenticating advertisements, or by designing the system so that the force multiplier is not effective, for example by requiring that messages in response to the advertisement are sent on a different channel to or at a lower priority than safety-of-life messages.
- *Privacy*: See extensive discussion above. To mitigate threats to *Users*, advertisements must either contain only common services or be authenticated. To mitigate threats to *Providers*, it should always be the case that offering a service is a conscious choice by the operator of the device.

Performance and Mechanisms:

- *Authenticity, integrity, authorization, non-repudiation*: Public key cryptography, digital signatures, and digital certificates using ECDSA. There is a bandwidth issue here. The certificate should state which services the Provider is entitled to offer. However, the Provider will not necessarily offer all services at all times. If the Provider has a certificate for the full set of services they are offering, this certificate might be very large. However, if the Provider has individual certificates for subsets of services, it may need a large (“exponential”) number of certificates, or it may run the risk of not having a certificate for a particular subset of services. The decision about how to address this is up to individual implementers.

4.5.1 Support in Standards

US:

- *Authenticity, integrity, authorization, non-repudiation, relevance*: Provided by IEEE 1609.2, which supports authenticated and unauthenticated services within the same advertisement.
- *Privilege classes*: Not currently specified in any standard.
- *Confidentiality*: Not needed.
- *Availability*: No standardized solution yet. This is likely to be addressed by individual implementers rather than by standards.

- *Privacy*: 1609.2 certificates for service advertisements are not anonymous. Although they may be anonymized by omitting the identifier, this would not support bulk revocation.
- *Anti-replay, performance*: 1609.2 allows for an advertisement to be signed once every few seconds rather than every time it is sent. The sender then resends the same advertisement until the next signature refresh time. The receiver does not have to verify duplicate advertisements. Replay is confirmation.

EU: No support yet.

4.6 Local High-Speed Unicast Service

Local high-speed unicast services are: unicast, single-hop, high time-criticality, medium data size, frequently advertised then used as needed, V2I/I2V, local sessions. Example: tolling

Threat model notes:

- *Advertisements*: These are location-based services and as such will need to be advertised: see the analysis in the previous section for specific security comments related to advertisements.
- *Trustworthiness of devices*: Unlike in previous sections, it is possible to assume that one side of the transaction (the toll plaza) is physically protected and is highly unlikely to be compromised, and has an online interaction with the central management service often enough to be useful. This enables the use of the performance benefits of symmetric cryptography.
- *Relationships*: A device may sign up for a service from a specific provider and then use that service multiple times. This means the device has an ongoing relationship with the provider and, so long as the provider is correctly identified, symmetric cryptography can be used, which greatly improves performance.

Services:

- *Confidentiality*: Required.
- *Authenticity*: Required.
- *Integrity*: Required.
- *Authorization and privilege classes*: Authorization is needed; privilege classes are application-specific.
- *Non-repudiation*: A transaction may be challenged, in which case there must be evidence to support the assertion that the transaction took place. In the case of tolling, this evidence is provided out-of-band, in general by cameras. In the case of other applications, cryptographic non-repudiation may be more practical.
- *Anti-replay and relevance*: Messages must be fresh; replay may be an attack.
- *Availability*: Application developers must have a backup plan for if the application becomes unavailable.

- *Privacy*: Privacy issues with the advertisement were addressed in the previous section. Privacy issues within the transaction can be addressed by encrypting the transaction.

Performance and Mechanisms:

- These applications often need to complete transactions very quickly, so minimizing packet size and processing time is very important. Symmetric cryptography is the natural mechanism to use if keys can be pre-shared between the device and the service provider.

4.6.1 Support in Standards

US: Tolling over 5.9 GHz is defined in IEEE 1609.11 (2011b). There is no standard for security for other types of local high-speed unicast service.

EU: Tolling over 5.8 GHz is defined in ISO 14096 (2004) and ISO 15628 (2003). There is no standard for security for other types of local high-speed unicast service.

4.7 Local Groupcast Service

Local groupcast services are as with local unicast service but groupcast. Example: Subscription information with a large subscribing group – for example, premium traffic information with high-quality recommendations for alternate routes.

The distinguishing features of this type of service are that (a) information is broadcast to the subscribers – it is in general non-interactive; (b) the service provider may want to provide the service to some but not all of the vehicles in a particular RSE comm zone, or in a particular larger region.

Threat model notes:

- *Advertisements*: As with local unicast service, these services are advertised. See the analysis of advertisements in 4.5.
- *Trustworthiness of devices*: As with local unicast service, it can be assumed that the service provider is physically protected, talks to central management as and when necessary, and can be trusted. The end devices are less trusted. A condition of sending the subscription information may be that the end devices do not share the content with non-subscribers.
- *Relationships*: A device has an ongoing relationship with the service provider. Symmetric cryptography can be used.
- *Groupcast*: The reason for using groupcast is to reduce bandwidth consumption: There is some per-subscriber overhead to get into the stream, but once a subscriber has access to the stream, only a single stream needs to be provided no matter how many subscribers are present. This approach means that different subscribers get the same data, and so it is difficult to tell which subscriber was responsible

for redistributing information. Groupcast is therefore a better approach for medium-value, time-sensitive content (such as sporting results) than for high value or persistent content (such as movies).

Services:

- *Confidentiality*: Required to prevent non-subscribers from getting subscriber-only information.
- *Authenticity*: Required.
- *Integrity*: Required.
- *Authorization and privilege classes*: Authorization is needed. Privilege classes may include the central service provider and one or more levels of subscriber.
- *Non-repudiation*: The service provider does not want the content to be shared. If the service provider discovers that content has been shared, it may be desirable to identify which subscriber was responsible.
- *Anti-replay and relevance*: Application-specific. Messages will in general need to be checked for freshness. Since the service is non-interactive, replay is repetition of previous information.
- *Availability*: By its nature this service is only available where the service provider has chosen to provide it. Lack of availability is not a significant threat to the system.
- *Privacy*: The service provider should ensure that the identity of subscribers is not revealed to third parties (other subscribers or non-subscribers) unless this is integral to the operation of the system. If this is integral to the operation of the system, there must be a mechanism where the subscriber can explicitly assent to their identity being shared and withdraw the assent if necessary.

Performance and Mechanisms:

- *Mechanisms*: The natural mechanism for the data distribution is to provide a symmetric key to all subscribers and use this to encrypt, authenticate, and integrity-check the subscription data. The service provider may want to support area- or time-period-specific keys, in which case there must be a mechanism to rollover keys. When a subscriber “connects to the stream” they authenticate to the server using either a public key mechanism or a pre-established symmetric key. When keys are rolled over, subscribers may need to reauthenticate. If rollover is desired, it may improve the robustness of the system to distribute keys slightly before the data protection key itself changes.
- *Traitor tracing*: If a service provider wants to determine which subscriber was responsible for sharing subscription information, they may need to include some “watermarking” mechanism for distinguishing between different subscribers’ data streams. If this mechanism is used, the system is no longer groupcast but a collection of unicast streams and does not fall in this category.

4.7.1 Support in Standards

Not currently supported in standards.

4.8 Low-Speed Unicast Service

Low-speed unicast services are: unicast, single-hop, low time-criticality, medium or large data size, low frequency, V2I/I2V with restricted local or remote session. A unicast service consumed at low speeds, such as fleet management or data calibration.

These services differ from high-speed unicast services in that the off-vehicle end of the communications session may be remote over the network. The application cannot rely on rapid exchange of large amounts of information and will have higher latency than the high-speed unicast service.

In general, these services will use an IP connection and so the use of existing IP security mechanisms may be appropriate.

Threat model notes: See high-speed unicast service.

Services:

- *Confidentiality:* Required.
- *Authenticity:* Required.
- *Integrity:* Required.
- *Authorization and privilege classes:* Authorization is needed; privilege classes are application-specific.
- *Non-repudiation:* Application-specific. There may need to be evidence to support an assertion that a transaction took place. This evidence may be provided out-of-band, such as by cameras, or by cryptographic non-repudiation may be more practical.
- *Anti-replay and relevance:* Messages must be fresh; replay may be an attack.
- *Availability:* Application developers must have a backup plan for if the application becomes unavailable for a long period.
- *Privacy:* Privacy issues with the advertisement were addressed in the previous section. Privacy issues within the transaction can be addressed by encrypting the transaction. If the client authenticates to the server, this must be done once an encrypted channel has been established to avoid leaking the client identity to an eavesdropper.

Performance and Mechanisms:

- *Cryptography:* The typical instantiation of a service of this type will involve the service provider and consumer authenticating to each other and establishing a bulk data key for the transaction (or agreeing to use an existing one). The authentication may be symmetric, by means of pre-shared keys, or public key. Symmetric authentication will be appropriate where the provider and consumer are already known to each other. Public key authentication will be appropriate if a consumer may come across many different service providers. For simplicity, if there will be a significant number of new service providers in the system, the system designer may choose always to use public key authentication, even for existing relationships.

4.8.1 Support in Standards

No ITS-specific standards yet in the US or EU. However, existing Internet services such as Internet Protocol Security (IPSec) (many references: see ([Wikipedia a](#)) for an overview) or Transport Layer Security (TLS)/Datagram Transport Layer Security (DTLS) (many references: see ([Wikipedia b](#)) for an overview) may be suitable. There has been work in the USA on an ITS-optimized version of TLS known as Vehicular DTLS (V-DTLS) but this is not yet standardized.

4.9 Distributed (Networked) Service

Distributed services are: unicast, single-hop, low time-criticality, medium or large data size, low frequency, V2I/I2V with local or persistent. Real-time traffic information, provisioning, maybe theft-related services.

This service is similar to the low-speed unicast service in that it involves connecting with a service provider across a network. However, the difference is that the logical communication session may need to persist across multiple “touches” between the OBE and an RSE offering access to the backhaul. The persistence may be provided at the application level, the transport layer, or the internet layer.

Threat model notes:

- *General:* The threat model considerations are similar to the considerations for high-speed unicast and low-speed unicast.
- *Privacy:* When a user reconnects to a session, the system should not reveal the user’s identity or a persistent session identifier to an eavesdropper. As noted above, if a service is connected to following an advertisement, the fact that the user is connecting to the service may be an identifier in itself. To reduce the information leaked by this, no other identifying information about the user should be included until a secured connection has been established.

Services:

- *Confidentiality:* Required.
- *Authenticity:* Required.
- *Integrity:* Required.
- *Authorization and privilege classes:* Authorization is needed; privilege classes are application-specific.
- *Non-repudiation:* Application-specific. There may need to be evidence to support an assertion that a transaction took place. This evidence may be provided out-of-band, such as by cameras, or by cryptographic non-repudiation may be more practical.
- *Anti-replay and relevance:* Messages must be fresh; replay may be an attack.
- *Availability:* Application developers must have a backup plan for if the application becomes unavailable for a long period.

- *Privacy*: Privacy issues with the advertisement were addressed in a previous section. Privacy issues within the transaction can be addressed by encrypting the transaction. If the client authenticates to the server, this must be done once an encrypted channel has been established to avoid leaking the client identity to an eavesdropper.

Performance and Mechanisms:

As with the other unicast application classes, there will be an authentication mechanism which may be public key or symmetric key, followed by a handoff to a symmetric cryptography-based bulk data protection protocol.

4.9.1 Support in Standards

There is no explicit support in ITS standards.

Within the IETF there has been and continues to be great interest in privacy for mobile IP. The problem statement is outlined in (Koodli 2007). Subsequent RFPs provide possible instantiations of solutions. There is no dominant solution in deployment.

4.10 Multiple Applications

An ITS-S may run multiple applications. Each application will have its own security requirements as described above. However, the combination of applications may introduce additional threats to the communications security, such as:

- Privacy – the combination of applications that an ITS-S runs may act as an identifier
- Availability – one application may consume resources needed by another application

These issues are mainly handled by mechanisms on the ITS-S before messages are transmitted. These mechanisms are described in ♦ Sect. 6. The one exception is privacy. If an ITS-S is transmitting application datagrams from multiple applications with the same network identifiers (such as the MAC address), an eavesdropper can tell that the applications are being run on the same platform. If the eavesdropper knows the identity of the sender (perhaps because they are legitimately participating in one of the applications), this is a leak of personal information; even if the eavesdropper does not know the identity of the sender, the combination of applications could be unique to the station and allow the eavesdropper to track the vehicle.

The natural countermeasure is to use a different set of network identifiers for each application, but this is difficult to implement down to the MAC level with off-the-shelf chipsets.

Standards support. There is no requirement in standards for different addresses for different applications. IEEE 1609.4 (2010b) supports changes of sending MAC address,

but to support different addresses for different applications on the same channel a device would have to receive on multiple MAC addresses simultaneously, and there is no standards support for this at the time of writing.

5 Security Management Services

The cryptographic security services described in this section rely on the authorization of cryptographic keys. For systems based on symmetric cryptography, where end-user devices use a symmetric key and an identifier to communicate with a central server, there must be a security management service that installs keys and identifiers on those end devices. For public key systems, which use digital certificates, there must be a security management service that allows devices to obtain certificates and allows the system to remove known misbehavers from the system.

The (U.S) National Institute of Standards and Technology identifies the following lifecycle stages in key management (Barker et al. 2007):

- Pre-operational:
 - User Registration – establish an identity within the security management system for the user or device.
 - System Initialization – establish algorithm preferences, trusted parties, and policies.
 - User Initialization – install and initialize cryptographic hardware or software.
 - Keying Material Installation – install keying material.
 - Key Establishment – generate and distribute, or agree, keying material for communication between entities. For public key cryptography-based systems, this can include distribution of locally or centrally generated asymmetric key pairs; distribution of static or ephemeral public keys; distribution of a trust anchor's public key in a PKI; certificate request by a Certificate Authority. For symmetric cryptography-based system, this can include local or central generation and distribution of keys. In both cases, the key establishment stage can include keys for key management as well as keys for application operations. (This stage is also known as the Bootstrap Trust stage).
 - Key Registration – in this stage, a key is bound to the appropriate entity or permissions. For a symmetric key system, this may be done by storing a symmetric key with the relevant attributes. For a public key system, this is certificate issuance.
- Operational:
 - Normal Operation – the key is stored securely by the entity that uses it, and accessed for use as appropriate.
 - Continuity of Operations – if the keying material is lost for some reason, it is recovered from secure backup. (It may be more appropriate in some systems simply to establish a new key or a whole new device).

- Key Change – rekeying with entirely fresh keying material using the key establishment and registration functions from the pre-operational phase; or updating the existing key in a non-reversible way, in which case it may be possible to maintain the binding without explicitly repeating the registration function.
- Key Derivation – if keys are derived from a secret value, derive a new key from that value.
- Post-Operational:
 - Archive Storage and Key Recovery – if access after operations is necessary.
 - Entity De-registration – Remove the authorizations of an entity to take some or all actions.
 - Key De-Registration – Remove the binding between a key and some attributes.
 - Key Destruction – Physical destruction of copies of keys.
 - Key Revocation – Remove a key from the system before the end of its normal lifetime.
- Destroyed: Key material is no longer available, though certain attributes may be preserved for audit purposes.

Not every system will go through each of these lifecycle stages, but any plan to deploy ITS units that use cryptographic must identify which stage will be used and specify procedures and technologies to implement that stage.

The rest of this section addresses key management lifecycle issues that are particularly relevant to ITS systems.

5.1 Initialization

ETSI TS 102 731 (2010) describes two stages in initializing the keying material for ITS stations. These stages assume that the ITS station has already been given material that allows it to trust authorities within the system, i.e., it has gone through the Bootstrap Trust stage.

- *Enrolment*: An ITS station registers with an enrolment authority (EA) and obtains an enrolment credential. The enrolment credential may be general (i.e., may cover all of the ITS-S's activities) or may be specific to an application. The enrolment credential is not used directly by applications, but is presented to authorization authorities to obtain authorization tickets.
- *Authorization*: The ITS station presents an enrolment credential to an authorization authority and obtains an authorization ticket. This ticket is a binding between key material and a set of attributes and is presented to other ITS-Ss during the course of application operations to authenticate and authorize transactions.

The above description is somewhat abstract but provides a framework to understand both symmetric and public key key management systems.

5.1.1 Public Key

Enrolment: The ITS station is initialized with a private key and applies for a long-term or “identity” certificate from a CA. This process must be carried out in some kind of physically secure environment, so that the CA has a guarantee that the ITS station is a valid station that is entitled to the certificate it is asking for. For example, the ITS station could be initialized on the factory floor, over an Ethernet connection that connects it to a VPN connection that is specific to the link between the CA and the factory. Or the CA could pre-generate pairs of private keys and certificates, where the certificates have the appropriate set of attributes, and distribute them to the factory which then directly installs them on the ITS-S. Whatever the specific protection processes are, it must be extremely difficult for an attacker to insert a non-valid device into the process and obtain a long-term certificate for it.

Secure initialization processes such as this are common in industry and manufacturers of devices such as smartcards, Subscriber Identity Modules (SIMs) for mobile phones, and Trusted Platform Modules that commonly use them to initialize keying material. The initialization process described above can potentially be carried out in two stages: In the first stage, a smartcard is initialized using existing processes, while in the second, it is physically inserted into the ITS-S and secured in such a way that if removed it will no longer work.

Authorization: The ITS station presents a certificate request to a CA, signed by the long-term certificate, and obtains a certificate with the appropriate permissions. It then uses this certificate to authorize application messages.

Other standards: The EA and AA play similar roles to the Long-Term Certificate Authority and Pseudonym CA in C2C-CC SEC ([Car 2 Car Communications Consortium n.d.](#)). IEEE 1609.2 does not discuss PKI structure but 1609.2 certificates may be used to support this architecture.

Implementation notes: At the time of writing there are no production CAs for 1609.2 certificates. X.509 certificate-based PKIs support this hierarchy.

5.1.2 Symmetric Key

Enrolment: The ITS-S communicates with a central key server and obtains a key management key or keys, known only to the server and the ITS-S. As with public key, this needs to be done in a physically secure environment so that the key server can be assured that the ITS-S is entitled to the keys it is requesting.

Authorization: The ITS-S communicates with the key server. The key server sends it a communications key, encrypted and authenticated using the key management keys. The communications key will frequently be derived from a key ID value using a communications master key: For example, the communications key could simply be

the key ID encrypted with the communications master key. The separation between communications keys and key management keys allows for there to be some entities in the system that communicate securely with all devices (because they have the communications master key) while other, more privileged entities can both communicate securely and instruct the devices to update the communications key (because they have the key management key).

Other standards: This distinction between key management keys and communications key is explicit in IEEE 1609.11, the standard for tolling.

Implementations: Many toll operators already implement this.

5.2 Key Update, Revocation, and Removal of Misbehaving Devices

If a device is malfunctioning, or if an attacker has obtained its keying material and is using that material to generate false messages, it must be removed from the system. “Removed” may mean a range of things, but in particular it means that the device must lose its ability to authorize itself to other devices, in other words that other devices must stop trusting the compromised device’s authorization tickets.

There are three basic ways to implement this.

- *Online checking:* when an ITS-S gets a message with an authorization ticket, it performs a real-time check with some server to see whether or not the ticket should be trusted. This is the equivalent of the Online Certificate Status Protocol (OCSP) in the PKIX world. The approach requires the ITS-S to be online to the server at all times. Unless network latency can be minimized, this approach is not suited to time-critical applications.
- *Ticket expiration:* Authorization tickets have an expiry time and cannot be used to authorize a message after that time. (In practice, since it may be useful to determine whether a message should have been trusted at the time of issuance, this is implemented by including a generation time in the message and checking that the generation time is within the validity period of the authorization ticket.) Each ITS-S must apply to the authorization authority for more ticket(s) before its current ticket(s) expires. The authorization authority will refuse to issue more tickets to a no-longer-valid ITS. This approach requires the ITS-S to have guaranteed bidirectional communications with the authorization authority from time to time.
- *Revocation:* information is distributed to possible recipients of messages from the compromised ITS-S, telling them that the ITS-S’s authorization tickets are not to be trusted. When an ITS-S receives an authenticated message it checks its local revocation database for information about the attached authorization ticket and rejects a message if the authorization ticket is revoked. This approach does not require the ITS-S to have bidirectional communications with the authorization authority; revocation information can be broadcast.

Network availability and latency make online checking unsuitable for the vehicle side of most applications, so the vehicular ITS-S will in general use ticket expiration and revocation to validate trustworthiness.

A prime consideration in deciding which of these mechanisms to use will be the channel congestion that each method might cause. The bandwidth requirements for ticket expiration depend on the ticket lifetime and the number of units in the system. The bandwidth requirements for revocation depend on the number of revoked units (which can be considered as total units times a revocation rate), the time that a revoked unit has to stay on the revocation list, and the number of times a revocation list needs to be sent to ensure that almost all units in the system have received it.

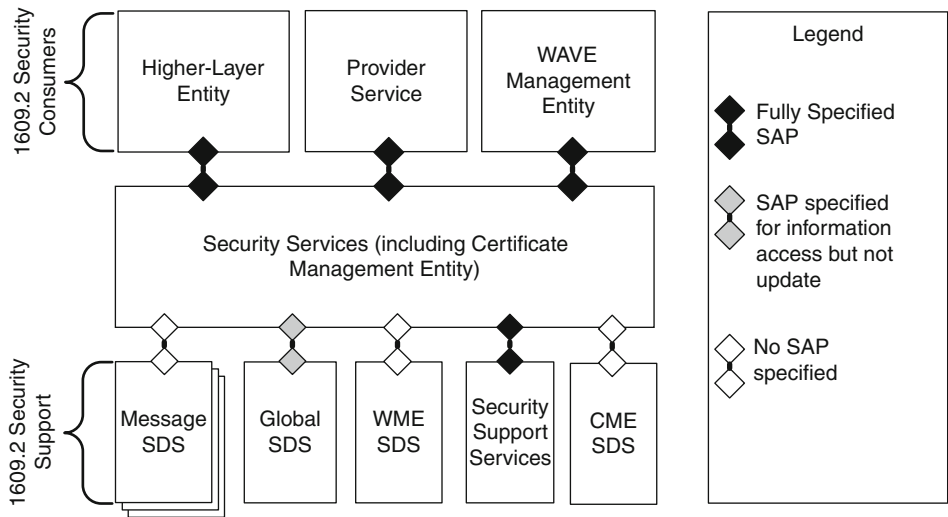
Boiling this down, the bandwidth requirements for an expiration-based system scale as N , the number of vehicles in the system, while the bandwidth requirements for revocation scale as N^2 (though possibly with a smaller multiplier). In practice a system will use a combination of the two. It makes sense to rely primarily on revocation lists if the system provides frequent one-way communications from the CA to the vehicles and very little two-way communications. In the early stages of ITS deployment, this may in fact prove to be the case: Technologies that are still in the research stage such as “epidemic distribution” (Laberteaux et al. 2008) effectively increase the one-way bandwidth from the CA in a situation where two-way bandwidth is still limited. However, as ITS penetration increases, two-way communications to the CA will become more available because (a) there will be more RSEs and (b) there will be a greater number of units with a non-5.9 persistent data connection.

5.3 Implementation and Support in Standards

Any implementation of an ITS-S application must consider the following questions:

- *Initialization:* Where does the implementation get its initial set of keys?
 - What authority provides the keys?
 - How are the keys put on the device or made available to the application in a secure way?
 - Over the air?
 - Over some wired connection?
 - What processes and procedures are followed?
- *Operations:*
 - How does the implementation update its own keys?
 - How does the implementation know to trust keys received from other parties?

Initialization: Initialization is the process where an ITS-S goes from a keyless state to a state where it has at least one set of keys that can be used for authorization and authentication to other ITS-S. No existing standards specify the procedures to be put in place to ensure initialization is secure, although some standards provide guidelines: 1609.2, C2C-CC.



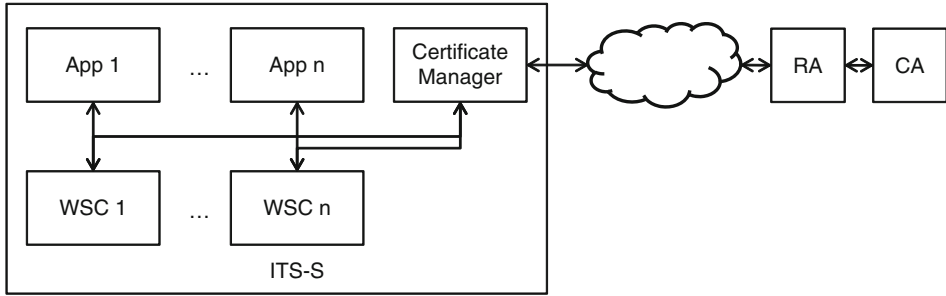
■ Fig. 49.1
1609.2 Architecture (Note: figure (c) IEEE)

Operations: In contrast to initialization, there exist descriptions of ongoing security management operations (once the first set of keys have been established) which are detailed enough to form the basis of implementations. Most descriptions of ITS stations in standards and other documents assume that security management is carried out by a different process from the application: 1609.2 describes a Certificate Management Entity; SeVeCom (Kung et al. 2007) assumes a Security Module. These descriptions are very similar to each other.

For example, 1609.2 provides a high-level architecture for the ITS-S, illustrated below, along with Service Access Points (SAPs) which allow an application to initiate certificate management processes. Each higher layer entity has its own Secure Data Store, which manages the keys and certificates used to sign outgoing messages on behalf of that entity, and stores the certificates of recipients who are trusted to receive encrypted messages from that sender. There is also a Global Secure Data Store (which is global in the sense that it is usable by all processes running on a particular ITS station) which stores lists of root certificates and revoked certificates and is available to the security processing in general. 1609.2 also specifies over-the-air messages for certificate request and CRL distribution in both a push and pull model, although it does not specify certificate request messages for anonymous applications (🔗 Fig. 49.1).

5.3.1 Device-Side Case Study: VIIC Proof of Concept

In the VIIC Proof of Concept, the OBEs were multi-application devices. Different applications were identified on the device by Provider Service Identifier (PSID),

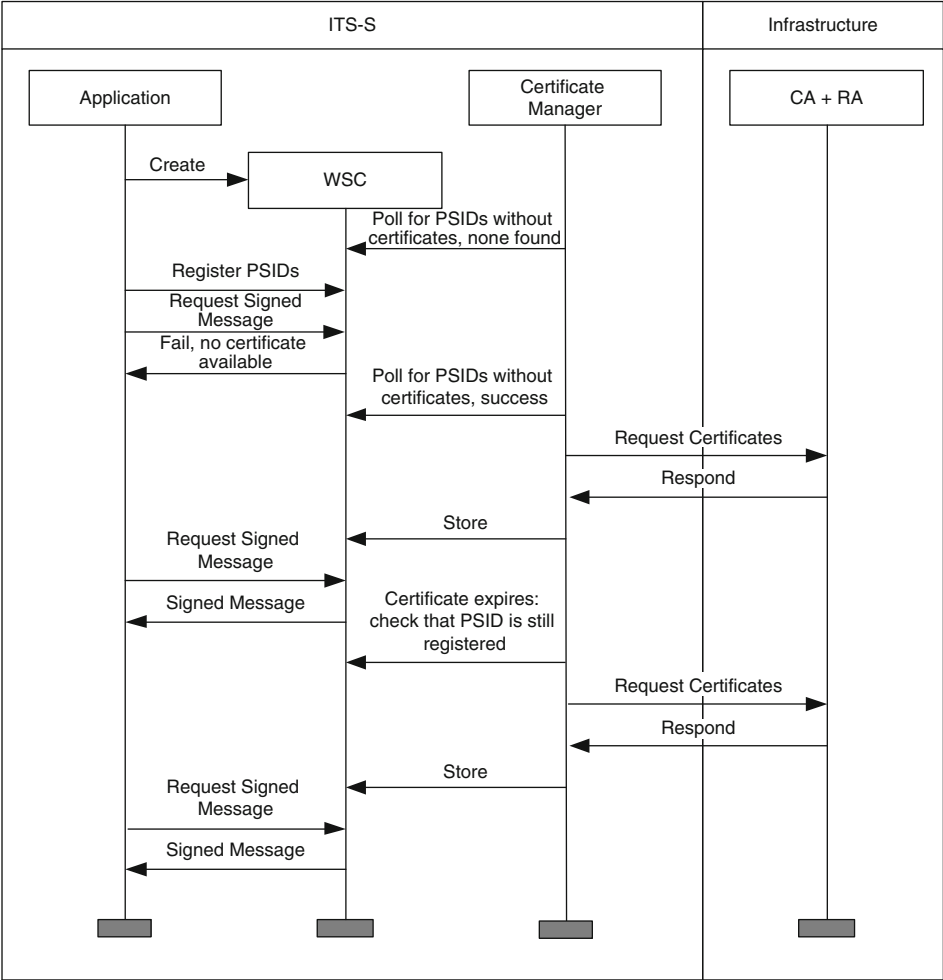


■ Fig. 49.2

ITS-S security architecture in VIIC PoC

which is the identifier also used in 1609.2 to specify a sender's permissions. Here, the architecture was as follows (Booz Allen Hamilton 2008; VII Consortium 2009) (● Figs. 49.2 and ● 49.3):

1. There was a single Certificate Authority in the system.
2. There was a single Certificate Manager on the device. This was a Java application.
3. On initialization, the Certificate Manager applied to the CA for a device certificate. This process was secured by manually approving each device certificate request and checking an ID in the request against the IDs of the approved devices (If there had been an attacker who knew a valid device ID and the CA public key, they could have used this to construct a device certificate request that would have had a good chance of being accepted as valid. This was considered acceptable for the field operational test as the chance of an attack was considered to be low, and the attack would have been discovered anyway when the valid possessor of that ID applied for a certificate. In practice, no attacks on this part of the process were observed).
4. Each application on the device, on registration, created its own Secure Data Store (in the terminology of 1609.2), known as a WAVE Security Context (WSC). The application registered its security attributes with the WSC. These include the PSID, whether the application is anonymous or identified, and so on.
5. The Certificate Manager periodically polled each WSC to see if the application had registered PSIDs for which there was no certificate, or the certificate was expired. For each such *uncertified PSID*, the Certificate Manager constructed a certificate request and sent it to the CA. When the response was received, the Certificate Manager wrote the resulting certificate into the application's WSC.
6. If an application did not have a valid current certificate, it did not send messages but instead waited for the Certificate Manager to complete receiving new certificates from the CA.
7. When the Certificate Manager was able to contact the CA, as well as requesting certificates for any not fully certified applications, it requested the latest version of CRLs for any registered applications.



■ Fig. 49.3
Certificate lifecycle from point of view of application in VIIC PoC

- 8. If an application’s certificate was revoked, the Certificate Manager detected this and, depending on an application-specific policy, would either apply for more certificates or stop applying for certificates, deactivating the application.
- 9. The system did not provide a means for deactivating an entire unit.

A key motivator for the VIIC PoC design was to reduce the burden on the application developer: All that the application needed to register with its WSC was information that the developer already knew, such as the PSID and whether or not the application was anonymous. This experience was a key to developing the concept of security profiles in 1609.2, which provide a means and opportunity for a standards development organization (SDO) to fully specify the security behavior of an application, simplifying the task for implementers.

5.4 Multiple CAs and Roaming

The previous discussion has focused on the case where there is a single CA. However, in full deployment, there may well be multiple CAs:

- Specific applications may use multiple CAs, just as there are multiple CAs issuing SSL certificates in the general Internet setting today, with browser providers taking responsibility for adding new CAs and removing CAs that have proved to be untrustworthy.
- There may be specific local requirements and policies that a CA has to fulfill, for example in individual countries. One country may not want to honor certificates issued by a CA in another country, or may simply require certificates issued by a domestic CA. In this case an ITS-S crossing the border will need to obtain certificates from a domestic CA in the new country.

ETSI uses the concept of Enrolment and Authorization domains to provide an abstract description of the multiple CA setting. In the ETSI terminology, an authority (Enrolment or Authorization) is authorized to issue credentials within a domain; once an ITS-S is no longer operating in that domain, it must seek authorization from a different authority. This covers the case when an ITS-S moves from one country to another, as well as the case where, for example, an e-payment application chooses to use a different payment service provider.

In the multi-CA environment the implementer must address the following issues:

- How does an application with the choice of several CAs know which CA to apply to?
- How does an application obtain the root certificates of all the CAs that might authorize messages the station will receive, or otherwise become able to trust those messages?
- If the application moves from one domain to another (in ETSI terminology), how does it get the contact details of the new CA to reauthorize with that CA?

At the time of writing there is no standard that provides mechanisms to address these issues. ETSI will specifically address issues in cross-border roaming.

5.5 Infrastructure-Side Case Study: National ITS Architecture (USA)

The National ITS Architecture Security document (Architecture Development Team 2007) distinguishes between two aspects of security in the US National ITS Architecture: *Securing ITS*, which refers to the services that secure information within the system, and *ITS Security Areas*, which refers to the ways in which the ITS system can improve the security of its users. A concept of operations for infrastructure-side services is given in (US Department of Transportation 2011). This document identifies 20 services, of which eight are related to information security. Of these eight security services, seven are

identified as essential features of the Core System and all are identified as not completely satisfied by existing systems. The services are grouped into eight subsystems, of which three (Misbehavior Identification, User Permissions, and User Security subsystems) are related to security.

The system recommends the use of 1609.2 certificates for communications over the dedicated ITS channel, and X.509/PKIX certificates for other communications channels. Both devices and applications will have to be certified. The certification process should not constrain the policy decisions of what needs to be certified.

There may be multiple Core Systems. They will need to share information about misbehavior. The certificates issued by one Core System must be acceptable by another Core System. (So this allows for the concept of roaming even within the USA.) There must be shared policies and procedures for revocation. Misbehavior reports can be sent from Center, Mobile and Field users to the Misbehavior Identification Subsystem. Credentials of misbehaving users may be revoked or may be allowed to expire. An Activity Diagram is defined for certificate request and revocation, and for suspicious behavior notification. The services of the Core System allow application developers to use the core system rather than develop individual approaches to security.

5.6 Privacy Against the CA

The previous discussion of privacy has mainly concerned privacy against eavesdroppers, which is provided by ensuring that the ITS-S has many certificates and changes the certificate it uses quite often. However, if the CA knows which user has which certificate, it can still track the user by observing signed messages and looking up the certificate owner. This allows an insider with access to the CA's databases to track a target without being detected.

This problem may be addressed in a number of ways.

- *Internal processes.* Assume the CA has internal processes that prevent this attack.
- *Separation of enrolment and authorization.* ETSI has a two-stage process for obtaining authorization tickets: First the ITS-S contacts an enrolment authority and obtains its enrolment credentials, then it presents the enrolment credentials to the authorization authority. (These are equivalent to the Long-Term CA and Pseudonym CA in the C2C-CC). The ITS-S canonical identity appears in interactions with the enrolment authority, and the enrolment credentials do not identify the ITS-S. This way the enrolment authority knows the identity of the vehicle but not what authorization tickets it has, and the authorization authority knows the authorization tickets but not the identity. However, the authorization authority still knows the full set of certificates issued to an individual vehicle and the location the vehicle was in at the time it made the certificate request (This is the network location, but this can typically be mapped to a physical location). This will need to be addressed with one of the other approaches described in this section. This approach is supported in standards by ETSI and by IEEE 1609.2.

- *Separation of request approval and issuance:* The PKI is separated into two functional entities, the Registration Authority (RA) and the Certificate Authority (CA). The ITS-S creates a certificate request, encrypts it for the CA, then signs it and sends it to the RA. The RA approves the request and sends it to the CA. The CA signs the certificates, encrypts them for the ITS-S, and returns them to the RA, which returns them to the ITS-S. This way the RA does not know the certificates, and the CA does not know the location or enrolment credentials of the ITS-S. This approach was used in the VIIC Proof of Concept project. At the time of writing there is no final European or American standard for preserving privacy with this approach, although there are research projects in America that aim to produce input to the IEEE standardization process.

6 Multi-application, Physical, Platform, and IVN Security

Previous sections in this chapter have focused on communications security. However, communications security depends on secure (meaning correctly behaving) endpoints. The practice of ensuring that endpoints are secure is known as platform security or systems security. This section discusses platform and systems security issues relevant to ITS stations, along with security issues on the in-vehicle network.

6.1 Vehicular Platforms

A vehicle contains a large number of control systems, most of which are implemented in a modern vehicle using electronic data communications (in contrast to older systems which were primarily mechanical). As a result, the vehicle contains a number of resources to which access must be limited by system security mechanisms. For example:

- Driving control systems – operating the brakes and the accelerator, possibly even steering
- System configuration – engine tuning, etc.
- Network resources – access to
- Application data – payment information for
- Access to personal or financial information
- Access to High Access Low Latency (HALL) channels which are intended primarily for use by safety-of-life applications
- Ability to turn on and off services used by other applications, such as GPS
- Access to cryptographic acceleration hardware
- Access to the human interface, including the ability to create believable hazard alerts

Additionally, the vehicle contains, generates, and uses information (such as telematics data and GPS position) that must be accurate.

There are two broad areas of concern here: correct operation of multiple applications on the ITS-S itself, and correct operation of the in-vehicle network.

ITS-S with multiple applications: Mechanisms to ensure the correct operation of a composite system such as found in a vehicle include:

- Different roles for users of the system, including typical user, administrator, security officer. Users acting in different roles have different abilities to install and run software, modify data, view or change cryptographic material, etc.
- Different user accounts, so that data available for read or write by one user is not necessarily available to another.
- Different applications, so that even if two applications are run by a single user the first application does not necessarily have programmatic access to the second application's data.
- A combination of trusted components and communications security.

Secure operations of the ITS-S itself are investigated within the EU OVERSEE project ([Oversee Project n.d.](#)). This project aims to define a secure framework with:

- Single point of access to vehicle networks
- Generic communication over multiple communication interfaces
- User specific rules for communication
- Protected runtime environments for the simultaneous and secure execution of applications (like Apps for the iPhone)
- Possibilities for platform and vehicle independent automotive applications (e.g., Open source projects)
- Open and standardized APIs

It is desirable that only authorized applications will be able to send at safety-of-life priority. Possibly in the future applications will have to undergo a certification process to demonstrate that they use only resources they are entitled to (US Department of Transportation 2011). The nature of this certification, however, has not yet been specified. At the time of writing, there is not even a vigorous requirements gathering process.

In-vehicle networking: The need to secure in-vehicle communications has been demonstrated by many high-profile attacks (Koscher et al. 2010). Preserving the security of in-vehicle network communications is made complicated by the fact that IVN components will tend to be very cost-sensitive. The EU EVITA project ([Evita Project n.d.](#)) has done a thorough analysis of this, building on previous research work (Wolf et al. 2004).

Architectures based on EVITA and OVERSEE are not mandated in any standards, but the projects can certainly be taken to outline a set of best practices that implementers would be wise to consider following.

6.2 Non-vehicular, Mobile, Multi-application Platforms

The infrastructure-side services must support many different types of mobile terminal, including built-in OBEs, retrofit devices, and personal aftermarket devices such as smartphones with ITS capabilities (US Department of Transportation 2011). Mobile

units may not even be in vehicles, but may be used by pedestrians or cyclists. There may be embedded vehicle terminals, aftermarket vehicle terminals, portable consumer electronic terminals, and infrastructure terminals. Policymakers will need to consider whether these different devices should have different restrictions on the messages they can send and resources they can consume. Implementers will need to work within the constraints set by policymakers. This policy making process is in the early stages at the time of writing but can be expected to develop rapidly as ITS approaches widespread deployment.

6.3 Security for the Long Term

Finally, as noted in Jentzsch et al. 2010, there is a further consideration that makes developing secure systems within a vehicle challenging: the long lifetime and challenges to upgrade of vehicular systems. A vehicle lifecycle is essentially 29 years: 4 years design, 5 years production, 20 years service. A design for protection of service interfaces in 2010 must still be effective in 2030. Key and certificate management systems must be effective over this time. Any system with a long lifetime that attempts to provide secure communications must be carefully designed to avoid errors.

Also, there is the possibility of unexpected advances in cryptanalysis, such as quantum computers which would destroy the security of ECDSA if developed (Perlnern and Cooper n. d.). The system must therefore not only be designed to current best practices but provide a smooth, secure upgrade path that can be used to protect against new security threats as they arise in the future. This is a challenge that the current standards do not yet meet. But they should.

7 Conclusion

Security designs for ITS are still at an early stage. Standards have laid some useful groundwork but there is still much work to be done in design, standardization, software architecture, and service provision. This chapter has provided an overview of design considerations and a snapshot of the current state of development. Future developers will have to rise to the challenge of fully realizing the system.

References

- 18 U.S.C. § 2721 et. seq. (1997) Drivers privacy protection act. Available at <http://uscode.house.gov/download/pls/18C123.txt>. Accessed 30 July 2011
- Adams JGU (1982) The efficacy of seat belt legislation? Transactions of the Society for Automotive Engineers, pp 2824–2838. Available at <http://john-adams.co.uk/wp-content/uploads/2006/SAEseatbelts.pdf>. Accessed 27 May 2011
- Architecture Development Team (2007) National ITS Architecture Security, U.S. Department of Transportation, Washington, DC
- Barker E, Barker W, Burr W (2007) NIST special publication SP 800-57, recommendation for key management – part 1: general. National Institute of Standards and Technology, pp 1–142. Available at <http://csrc.nist.gov/publications/nistp>

- ubs/800-57/sp800-57-Part1-revised2_Mar08-2007.pdf. Accessed 30 July 2011
- Booz Allen Hamilton (2008) Vehicle Infrastructure Integration (VII) proof of concept (POC) test final report (Executive summary)
- Buttyán L et al (2009) SLOW: a practical pseudonym changing scheme for location privacy in VANETs. In: IEEE VNC. Tokyo
- Car 2 Car Communications Consortium (n.d.), C2C security working group CAM/DENM security summary
- Cohen A, Einav L (2003) The effects of mandatory seat belt laws on driving behavior and traffic fatalities. *Rev Econ Stat* 85(4):828–843, <http://www.mitpressjournals.org/doi/abs/10.1162/003465303772815754>
- Deutschland T (2009) Sichere Intelligente Mobilität Testfeld Deutschland Kommunikationsprotokolle
- ETSI (2010) ETSI TS 102 731: Intelligent Transport Systems (ITS); Security; Security services and architecture
- ETSI (2011) ETSI TS 102 867: Intelligent Transport Systems (ITS); Security; Stage 3 mapping for IEEE 1609.2; IEEE 1609.2 profile
- European Data Protection Supervisor (2010) Opinion of the European data protection supervisor on the communication from the commission on an action plan for the deployment of intelligent transport systems in Europe and the accompanying proposal for a Directive of the European Parliament and of the Official Journal of the European Union, pp 47/6–47/15. Available at <http://dialnet.unirioja.es/servlet/articulo?codigo=2156793>. Accessed 24 May 2011
- Evita Project (n.d.) EVITA. Available at <http://evita-project.org/>. Accessed 9 July 2011
- Gerlach M, Guttler F (2007) Privacy in VANETs using changing pseudonyms-ideal and real. In: Vehicular technology conference 2007 VTC2007Spring IEEE 65th. pp 2521–2525. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4212947
- IEEE Vehicular Technology Society (2006) IEEE 1609.2-2006, Trial-use standard for wireless access in vehicular environments— security services for applications and management messages. IEEE Intelligent Transportation Standards Committee, Piscataway
- IEEE Vehicular Technology Society (2010a) IEEE Std 1609.3-2010, Standard for wireless access in vehicular environments (WAVE) – networking services
- IEEE Vehicular Technology Society (2010b) IEEE Std 1609.4-2010, IEEE Standard for wireless access in vehicular environments (WAVE) – multi-channel operation
- IEEE Vehicular Technology Society (2011a) IEEE 1609.2-2011, Standard for wireless access in vehicular environments – security services for applications and management messages, IEEE, Piscataway, NJ
- IEEE Vehicular Technology Society (2011b) IEEE td 1609.11-2010: IEEE standard for wireless access in vehicular environments (WAVE) – over-the-air electronic payment data exchange protocol for intelligent transportation systems (ITS), IEEE
- International Standards Organization (2003) ISO 15628 Road transport and traffic telematics – dedicated short range communication (DSRC)—DSRC application layer
- International Standards Organization (2004) ISO 14906 Road transport and traffic telematics – electronic fee collection – application interface definition for dedicated short-range communication
- Jentzsch A, Hackstein B, Goß S (2010) Security under automotive conditions and its influence on the product development process. In: Embedded world conference class 1.5, cryptography and embedded security. Available at http://www.embedded-world.eu/program/day-1.html?program_id=2310
- Johnson D, Menezes A (n.d.) The elliptic curve digital signature algorithm (ECDSA), Waterloo
- Koodli R (2007) RFC 4882: IP address location privacy and mobile IPV6: problem statement. Available at <http://www.ietf.org/rfc/rfc4882.txt>. Accessed 3 June 2011
- Koscher K et al (2010) Experimental security analysis of a modern automobile. In: 2010 IEEE symposium on security and privacy. IEEE, pp 447–462. Available at <http://www.computer.org/portal/web/csdl/doi/10.1109/SP.2010.34>. Accessed 9 July 2011
- Kung A et al., with SeVeCom (2007) SeVeCom deliverable 2.1 – security architecture and mechanisms for V2V/V2I. Available at http://www.sevecom.org/Deliverables/Sevecom_Deliverable_D2.1_v3.0.pdf. Accessed 30 July 2011
- Laberteaux KP, Haas JJ, Hu Y-C (2008) Security certificate revocation list distribution for vanet.

- In: Proceedings of the fifth ACM international workshop on VehiculAr Inter-NETworking – VANET’08, 88 p. Available at <http://portal.acm.org/citation.cfm?doid=1410043.1410063>
- National VII Coalition (2007) Vehicle infrastructure integration privacy policies framework, version 1.0.2
- Oversee Project (n.d.) OVERSEE. Available at <https://www.oversee-project.com/>. Accessed 9 July 2011
- Perlner RA, Cooper DA (n.d.) Quantum resistant public key cryptography: a survey. Quantum. National Institute for Standards and Technology
- Preciosa Project (n.d.) PRECIOSA – privacy enabled capability in co-operative systems and safety applications. Available at <http://www.preciosa-project.org/>. Accessed 11 July 2011
- Preserve Project (n.d.) www.preserve-project.eu, Preparing secure V2X communication systems. Available at <http://www.preserve-project.eu>. Accessed 11 July 2011
- RITA Joint Programs Office (2011) Intelligent Transportation Systems (ITS) standards program strategic plan for 2011–2014, Available at http://www.its.dot.gov/standards_strategic_plan/stds_strat_plan.pdf. Accessed 30 July 2011
- Schneier B (1996) Applied cryptography, Wiley. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.2838&rep=rep1&type=pdf>. Accessed 30 July 2011
- SeVeCom (2006) SeVeCom deliverable 1.1 – VANETS security requirements final version, Available at http://www.sevecom.org/Deliverables/Sevecom_Deliverable_D2.1_v3.0.pdf. Accessed 30 July 2011
- Society for Automotive Engineers (2009) SAE J2735, Dedicated short range communications (DSRC) message set dictionary, Available at http://standards.sae.org/j2735_200911. Accessed 30 July 2011
- Society for Automotive Engineers (n.d.) SAE J2945, Dedicated short range communication (DSRC) minimum performance requirements. Available at <http://standards.sae.org/wip/j2945>
- US Department of Transportation (2011) Core system concept of operations (ConOps), Revision C
- Vehicle Safety Communications Consortium (2005) VSCC final report appendix H : WAVE/DSRC security
- Vehicle Safety Communications Applications (n.d.) Vehicle safety communications applications (VSC-A) task 5 interim report II vol IV: vehicle security for communication-based safety applications
- VII Consortium (2009) Final report : vehicle infrastructure integration proof of concept executive summary – vehicle. Security. Available at http://ntl.bts.gov/lib/31000/31000/31079/14443_files/14443.pdf
- VIIC (2008) Vehicle infrastructure integration (VII) final report. VII Consortium, Novi, MI
- Wikipedia a. Wikipedia: IPsec. Available at <http://en.wikipedia.org/wiki/IPsec>. Accessed 3 June 2011
- Wikipedia b. Wikipedia: transport layer security. Available at http://en.wikipedia.org/wiki/Transport_Layer_Security. Accessed 3 June 2011
- Wolf M, Weimerskirch A, Paar C (2004) Security in automotive bus systems. In: Workshop on embedded IT-security in cars. Citeseer, pp 1–13. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.728&rep=rep1&type=pdf>. Accessed 14 Nov 2010

Section 10

Fully Autonomous Driving

Christian Laugier

50 Autonomous Driving: Context and State-of-the- Art

Javier Ibañez-Guzmán¹ · Christian Laugier² · John-David Yoder³ ·
Sebastian Thrun⁴

¹Multimedia and Driving Assistance Systems, Renault S.A.S,
Guyancourt, France

²e-Motion Project-Team, INRIA Grenoble Rhône-Alpes, Saint
Ismier, Cedex, France

³Mechanical Engineering Department, Ohio Northern University,
Ada, OH, USA

⁴Stanford University, Stanford, CA, USA

1	<i>Introduction</i>	1273
2	<i>Societal, Technological and Economical Motivators</i>	1275
3	<i>Vehicle Intelligence and the Navigation Functions</i>	1278
3.1	Interaction	1285
4	<i>Classification of Technological Advances in Vehicle Technology</i>	1286
4.1	Driver Centric	1287
4.2	Network Centric	1287
4.3	Vehicle Centric	1288
4.4	Evolution of Vehicle Architectures	1288
5	<i>Driver Centric Technologies</i>	1288
6	<i>Network Centric Technologies</i>	1293
7	<i>Vehicle Centric Technologies</i>	1296
8	<i>State-of-the-Art in Fully Autonomous Driving Research</i>	1299
8.1	Early US Unmanned Vehicle Projects	1300
8.2	Early EU Projects on Intelligent Vehicles	1301

8.3 US DARPA Challenges on Autonomous Driving 1302

8.4 EU Cybercars Projects 1303

8.5 Lessons Learned from DARPA Challenges 1304

9 Discussion, Conclusions and Structure of the Chapters 1305

Abstract: Vehicles are evolving into autonomous mobile-connected platforms. The rationale resides on the political and economic will towards a sustainable environment as well as advances in information and communication technologies that are rapidly being introduced into modern passenger vehicles. From a user perspective, safety and convenience are always a major concern. Further, new vehicles should enable people to drive that presently can not as well as to facilitate the continued mobility of the aging population.

Advances are led by endeavors from vehicle manufacturers, the military and academia and development of sensors applicable to ground vehicles. Initially, the motivators are detailed on the reasons that vehicles are being built with intelligent capabilities. An outline of the navigation problem is presented to provide an understanding of the functions needed for a vehicle to navigate autonomously. In order to provide an overall perspective on how technology is converging towards vehicles with autonomous capabilities, advances have been classified into driver centric, network centric and vehicle centric. Vehicle manufacturers are introducing at a rapid pace Advanced Driving Assistance Systems; these are considered as Driver Centric with all functions facilitating driver awareness. This has resulted on the introduction of perception sensors utilizable in traffic situations and technologies that are advancing from simple (targeted to inform drivers) towards the control of the vehicle. The introduction of wireless links onboard vehicles should enable the sharing of information and thus enlarge the situational awareness of drivers as the perceived area is enlarged. Network Centric vehicles provide the means to perceive areas that vehicle onboard sensors alone can not observe and thus grant functions that allow for the deployment of vehicles with autonomous capabilities. Finally, vehicle centric functions are examined; these apply directly to the deployment of autonomous vehicles. Efforts in this realm are not new and thus fundamental work in this area is included. Sensors capable to detect objects in the road network are identified as dictating the pace of developments.

The availability of intelligent sensors, advanced digital maps, and wireless communications technologies together with the availability of electric vehicles should allow for deployment on public streets without any environment modification. Likely, there will first be self-driving cars followed by environment modifications to facilitate their deployment.

1 Introduction

Ground transportation systems have evolved from simple electromechanical systems to complex computer-controlled networked electromechanical systems. Vehicles were initially used for leisure purposes, but they have become an integral part of our daily lives as very convenient forms of transport. As a result, road infrastructure has been built over the years and the manufacture of large volumes of vehicles has led to a reduction in their cost and their affordability by the public at large. However, these changes have led to different

problems such as traffic accidents resulting in death or injury, traffic jams, pollution, fierce competition in terms of cost, the depletion of fossil fuel reserves, etc.

Within the past years society has changed its perception with regard to transportation systems. Vehicles were regarded as a source of convenience and social status that provided industry much freedom, but today transportation systems are a source of increasing concern due to the high numbers of accidents, ecological constraints, high fuel costs, etc. involved. Governments, industry and society in general are moving to what is known as sustainable means of transport, to address the referred issues.

Vehicle manufacturers are obliged to constantly modify their offerings, to look for different solutions that make their products more innovative while they take into account the demands of society in terms of safety, pollution reduction, and connected mobility. Technological advances in information and communications technologies (ICT) have opened opportunities for the introduction of new functions onboard current vehicles; these are equipped by different types of proprioceptive sensors and there is a gradual introduction of exteroceptive sensors, like video cameras, radars, etc. Further, different propulsion systems are being used and most vehicles have their own computer networks with different nodes controlling different vehicle functions. In addition, vehicles can form networks that include other vehicles and the infrastructure bringing permanent connectivity to them. Thus there is a gradual transformation of the automobile from a single, largely mechanical entity into advanced intelligent interconnected platforms (Mitchell et al. 2010).

Within this context, the automotive industry is undergoing a transformation while robotics research in academia moves from indoor mobile platforms to the use full-scale vehicles operating with advanced levels of autonomy. New vehicles are using different propulsion systems that make them more computer-controllable, include multiple sensors for vehicle navigation functions, and are becoming nodes as part of large communications networks.

In this section the state-of-the-art in Intelligent Vehicles will be presented from a vehicle navigation perspective as these achieve autonomous navigation capabilities. The section is structured as follows:

1. *Motivation*: The motivation leading to this ongoing transformation of modern vehicles are presented in terms of usage, safety and external factors such as fossil-fuel constraints, pollution.
2. *Vehicle navigation functions*: The state-of-the art review is formulated in terms of vehicle navigation functions to focus the section on the machine intelligence and decision-making processes that are being developed and introduced to transform modern vehicles into connected platforms with autonomous navigation capabilities. It addresses the issues of autonomy, driver needs and communications. That is, it formulates the vehicle onboard intelligence as a navigation problem and thus defines the functional needs for vehicles to demonstrate autonomous navigation capabilities.
3. *Related vehicle technologies*: Current developments have been classified under three different perspectives: (1) *Driver Centric* addresses systems that seek to increase the

situational awareness of drivers providing different types of driving assistance systems that include the driver in the decision making process. (2) *Network Centric* addresses the use of wireless networks enabling the sharing of information among vehicles and the infrastructure creating awareness of the drivers and machines beyond what standalone vehicle systems could observe. (3) *Vehicle Centric* addresses systems that seek to convert vehicles into fully autonomous vehicles, with the driver outside the control loop. These different perspectives will be defined, presenting current developments in academia and industry.

4. *Future developments*: A perspective on future developments and on how these technologies could be adopted taking into account cost, legal and societal constraints will be provided.

2 Societal, Technological and Economical Motivators

In order to understand the rationale behind developments made on Intelligent Vehicles, it is important to have an introductory background of the enablers that have pushed and pulled the technological developments found today in current vehicles and advanced prototypes. These can be divided into those that encompass the Market and Customers, the Environment and Industry, Business Trends and Threats, and Strategy associated to the industry in general, as shown in Fig. 50.1.

Market and Customers. As a result of the globalisation process and the ability to remain connected anywhere in the world via the Internet and associated technologies, there is a growing demand for *Mobility*. This is identified as playing a vital role in the development of any modern society. Within this context vehicles are part of a larger system where multi-modal transportation systems co-exist in a sustainable manner. *Connectivity* is what has changed the working practices and daily lives of modern society. To see this, it is sufficient to observe the proliferation of mobile platforms that maintain the workforce in the industrialized world or the deployment of mobile telephones in emergent economies.

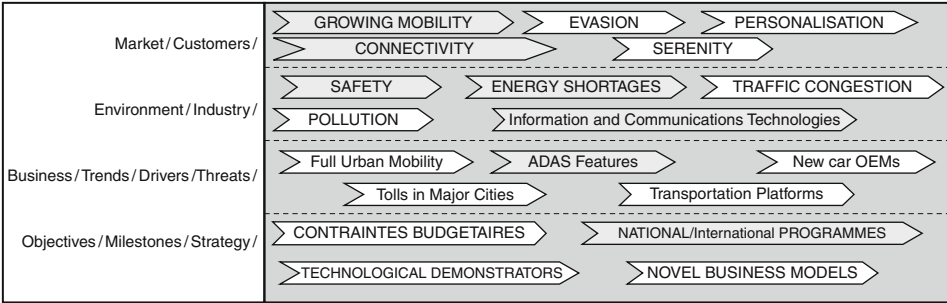


Fig. 50.1 Rationale influencing the development of vehicle technologies

Thus once the public is embarked in ground transportation systems, this connectivity needs to be maintained and perhaps extend its usage to the mobile platform themselves. Other motivators include the need to provide personalized platforms, to reduce the driving stress of drivers, in particular those of an aging population.

Environment and Industry. There are three major sources of concern influencing today's automotive industry: ecological, limited fossil fuels and safety; these are virtually transforming transportation systems. Concerns on the *limits to growth* are very much documented in the literature, these imply that *Pollution* is reaching in many cases critical proportions; *Traffic Congestion* has become almost intolerable, *Energy Shortages* are predicted, with an exponential growth on the use of vehicles in emergent economies. This has led to an increased interest on Intelligent Transportation Systems.

Despite much progress *accident statistics* in advanced economies remain high while those in emergent countries are reaching epidemic proportions. The burden of road crashes costs 1–3% of the world's GDP (IRAP 2007). Over the last 30 years overall *road safety* has been improving in industrialized countries, showing that political willingness and the application of countermeasures produce results. These included strong law enforcement, road infrastructure, and the introduction of different safety mechanisms (TRACE 2010). By contrast the opposite is occurring in rapidly developing economies.

Initial efforts at improving automotive safety centred on the deployment of passive safety systems such as shock absorbing chassis, safety belts, airbags, etc. This is accident mitigation – reducing accident damages using systems that were triggered once a collision occurred. Current efforts aim at intervening or warning drivers before an accident occurs, the earlier the intervention is, the more time systems will have to react. Accident statistics (accidentology) enable the targeting of applications (a thorough analysis of the conditions on which certain types of accidents occur) and thus are used for the design of safety-related driving assistance systems. They have pointed out that driver error is by far (95%) the most common factor implicated in vehicle accidents (followed by road/weather condition 2.5%, mechanical failure 2.5%) (NHTSA 2008). For example, in Europe, two thirds of fatal accidents were in rural roads with only 7% in motorways; pedestrian fatalities represent 33% in urban areas and only 9% in rural areas (2007). Active safety systems should address such situations. Understanding the conditions under which accidents occur should lead to the design of the most appropriate measures. It was found, for example, that road intersection accidents in Europe represented 21% of all fatalities, with a high incidence for the elderly. Other more punctual information was outlined in (Ibanez-Guzman et al. 2010) where major conditions of accidents at road intersections occurred were identified, as summarised in ► [Table 50.1](#). The use of such information leads to the type of complexity that the onboard machine intelligence needs to address if an autonomous vehicle is to traverse safely a road intersection. Vehicles are social entities that share a road network, thus the causes of accidents by other vehicles are important as these should be addressed by the autonomous platform if it is to be used in real traffic conditions.

The rapid growth in *Information and Communications Technologies* (ICT) is rapidly spreading onto vehicle platforms; Moore's law perfectly applies to an increased use of

■ Table 50.1

Context in which most road intersection accidents occur (Ibanez-Guzman et al. 2010)

Item	Description	Data
1.	Road geometry	Roads perpendicular to each other (53%). “Cutting Edge” situations represent 25% of accidents
2.	Type of regulation	Intersections with traffic lights (45–68%)
3.	Environment	Rural predominant EU-27 64% fatalities most countries
4.	Light and visibility conditions	Daylight & Twilight (67–75%)
5.	Weather conditions	Normal (82–90%); road surface Dry
6.	Main actors	2- Vehicles (67–82%); 1-pedestrian (9–14%); Passenger Vehicles, followed by motorcycles and pedestrians
7.	Driver age	The Elderly (37% of fatalities)

computer power as part of safety critical systems, multimedia centres, etc. Different types of computer processors are rapidly being incorporated as part of modern vehicles. Further, vehicles are equipped by a series of microcontrollers, networked using dedicated buses like the CAN-bus, MOST (media) or Flexray (safety). Advances in wireless communications from basic links at 440 MHz and low transmission rates to the use of 4 G links, have shown that Edholm’s law on the convergence between wired and wireless links is being applied (Webb 2007). Today frequencies have been allocated exclusively for the use of vehicle-related applications as for the case of Cooperative Vehicles in Europe. ICT onboard vehicles are simply following empiric laws used in the computer industry to predict progress.

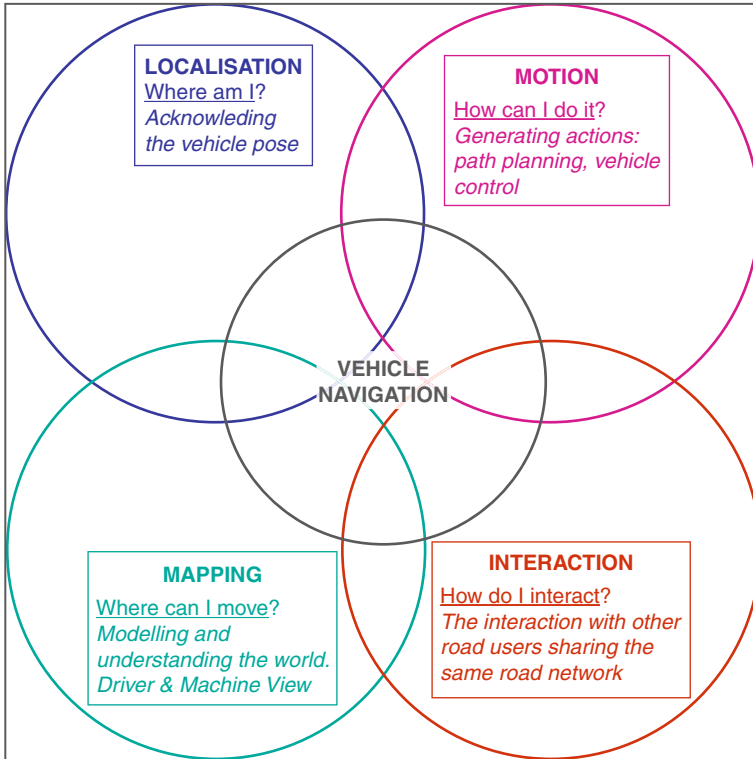
Business, Trends and Threats. Changes in the manner people live and work are transforming the transport industry. From the 1800s to the 1970s, populations created large metropolitan areas like London and New York. Between the 1970s and the 2010s these have led to the creation of Mega Cities like Shanghai and Mexico City. It is predicted that between the ~2010s and ~2025s the trend is towards a larger urban expansion or what is known as Urban Morphology to form areas like the Greater Paris, Greater Moscow. The new paradigm being the creation of Smart Cities where Energy, City Planning and ICT should lead to novel interconnected transport systems, and the need for *Full Urban Mobility*. Competition in the car industry is fierce, market needs have evolved, and the ownership of a vehicle is determined not only by being a source of transport but by providing safety, reliability and additional functions that make the offer more attractive. Within this concept *Advanced Driving Assistance Systems* (ADAS) and *Connectivity* are the functions providing differentiation, particularly for middle price range cars where cost is a major factor. Fundamental changes in the DNA of automobiles in terms of powertrain might lead to the emergence of *new manufacturers* of electric vehicles that could revolutionize the industry and represent a threat to established companies.

Strategy and Milestones. Governments and institutions like the European Commission (EC) are implementing novel policies to address these challenges. The EC for example has issued a directive to encourage the deployment of ITS across the EU with long-term implications towards the use of safety functions (European Parliament 2010). In Japan the “Innovation 25” plan seeks to create an ecological, interactive and secure society by 2025, where ITS plays a most important role for example, the Advanced Highway System (AHS and Advanced Safety Vehicle (ASV) (Lanson and Dauchez 2010)). Today, an integrated and systematic approach is being applied, it is inspiring industry and research. For example in Europe, initiatives towards a systematic integrated safety approach have been encouraged by the consortium of vehicle manufacturers (EURCAR 2010). It has led to accident studies that are providing a better understanding of the context under which accidents occur, and to the deployment of sensor-based safety systems onboard different types of vehicles as per the EU projects PREVENT, AIDE, etc. (EURCAR 2010). Successful demonstrations on the use of communications technologies for cooperative safety applications like the SKY project in Japan (Fukushima 2010) and EC sponsored projects like SAFESPOT, CVIS and Coopers have shown the advantages of using wireless communications technologies for enhancing safety (SAFESPOT 2010).

3 Vehicle Intelligence and the Navigation Functions

The fundamental function of a vehicle is to provide the ability to the driver, passengers or goods to move from a starting point to a finishing point in a safe and optimal manner. From the driver’s perspective and the use of any level of computer controlled functions (e.g., for comfort, safety, and networking), vehicle navigation functions could be characterised as consisting of four basic functions: Mapping, Localisation, Motion, and Interaction. These answer four basic navigation questions: *Where am I? Where I can move? How can I do it? and How do I interact?* If a vehicle is to navigate as expected, these functions need to operate correctly; these can be represented as the intersection of these functions in ● Fig. 50.2. If any of these functions performs poorly, the vehicle will not navigate as expected. Vehicles traverse road networks that are shared by other entities like pedestrians, and other vehicles that are expected to obey a set of pre-agreed traffic rules. Road networks are environments where unpredictable events might occur, where different actors have different levels of driving skills, and where human errors are the cause of most accidents. If the cause of an accident is due to motion without driver intervention, then the vehicle manufacturer might be liable, by contrast if the driver remains in the control loop the liability will be with the driver. This is very important in terms of reliability, safety and overall system integrity; it reflects the reluctance shown on the automation of certain driving tasks by vehicle OEMs.

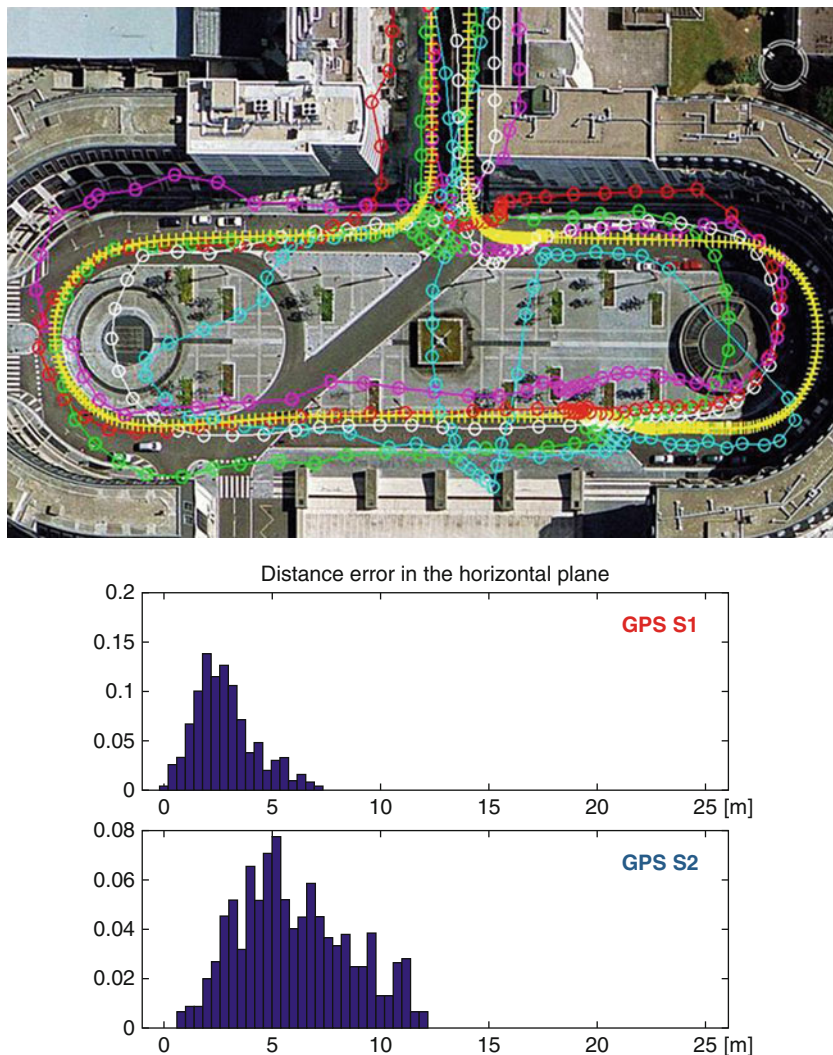
Localisation. This function can be defined as knowing the vehicle pose (position and orientation) with respect to an absolute or relative reference coordinate frame or determining the whereabouts of the vehicle. Global coordinates such as those used in Global Navigation Satellite Systems (GNSS) like GPS provide absolute information. The current



■ Fig. 50.2

Fundamental navigation functions

standard is the World Geodetic System (WGS 84) that establishes its origin at the Earth's centre of mass and the meridian zero longitude next to the Greenwich Prime Meridian (WGS84 2011). The relative location of a vehicle could be expressed with respect to a frame at a road intersection, or with respect to other vehicles. Absolute location estimations rely on weak radio signals from constellations of GNSS satellites which can be easily occluded and subject to errors due to noise and disturbances (Le Marchand et al. 2008). To compensate for these errors different dead reckoning, fault detection and fusion algorithms are used to estimate the vehicle location by fusing data from GPS receivers with data from exteroceptive and proprioceptive sensors. While good solutions exist, these rely on costly equipment like navigation level Inertial Navigation Units (IMU) or external RF corrections like RTK, and thus deployable solutions on passenger vehicles remain a challenge (Skog and Handel 2009). The difficulties encountered in using only GPS receivers in urban environments is illustrated by Fig. 50.3a. The large location estimation errors for a collection of receiver tests shown was identified as being caused by the multipath resulting from the surrounding buildings in the test scene (Le Marchand et al. 2009). The biases introduced by multipath cause estimation errors that persist over long periods. Figure 50.3b shows the strong differences that could exist between



■ Fig. 50.3

(a) Automotive type GPS responses. (b) Error spread of GPS receivers operating under the same conditions. Responses of automotive type GPS receivers in urban environments quantitative evaluation


GNSS receivers and the error spread to which localisation systems are subject. The quantitative evaluation was made by simultaneously recording the performance from different solutions with respect to a ground truth.

The “Where am I?” question is represents a fundamental requirement for vehicle navigation. Knowing the coordinates of a vehicle position is insufficient for vehicle navigation. It is necessary to know the context that is to project location information onto a map that will provide the vehicle driver or vehicle intelligence the ability to relate

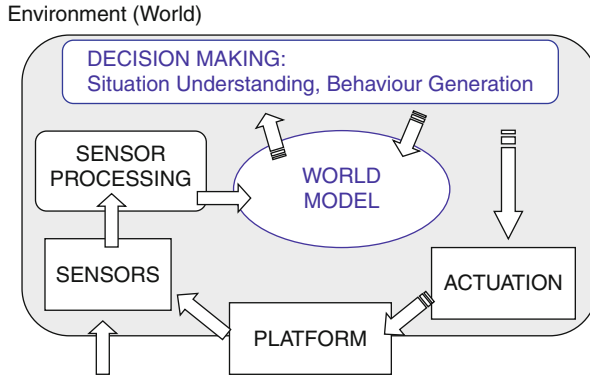
the whereabouts of the vehicle in the road network. Currently digital maps are commercially available and used extensively in onboard vehicle navigation systems; major suppliers include Navteq and TeleAtlas (TomTom). These maps hold road geometries and attributes associated to the links and nodes representing the road network. The projection of the vehicle location into the digital maps uses map-matching techniques that take into account errors on the location estimates as well as on the digital maps (Quddus et al. 2007; Fouque and Bonnifait 2009).

The location of a vehicle for autonomous navigation in all conditions is a challenge, as this is dependent on the absolute location estimates, the quality of the digital maps and the map-matching algorithms. For autonomous vehicle guidance it is thus not only a localisation problem but is also on determining with certainty the context where the vehicle evolves. Digital models of the environments are fundamental for autonomous vehicle navigation.

Vehicles can be localised at the same time as building maps that are used for navigation purposes by using the Simultaneous Localisation and Map Building (SLAM) approach that minimises or bypasses the need for GNSS signals (Guivant and Nebot 2001). New approaches consider maps as probability distributions over environment properties rather than fixed representations of the environment at a snapshot in time. The environment is modelled as a probabilistic grid, instead of a spatial grid, approach that allows for the reduction in uncertainty. By storing maps as probability models and not just expected values, it is possible to describe any environment much better (Levinson and Thrun 2010). Contrary to the use navigation digital maps that have been built for driver guidance, navigation maps that are built concurrently as the vehicle localizes itself, represent the likely path that the vehicles will follow and thus are closer to the expected vehicle position.

Mapping. Vehicle guidance entails understanding the spatio-temporal relationship between the vehicle and its environment. This can be regarded as modelling and understanding the world from the driver or computer controller perspective. Human or Machine Perception provide information about the vicinity of the vehicle environment. This is stored and represented in an abstract map and used at different granularity levels. The environment model is then used to gain understanding of the vehicle relationship with its environment and decide “Where can I move?” A driver perceives the environment which is to be traveled and builds a mental model using stored information of similar situations will be incorporated as well as subconscious recalls to driving legislation. These are used to gain an understanding of the situation and to decide what action to take to safely manoeuvre the vehicle. A driving assistance system or autonomous navigation system will follow a similar process (Langer et al. 1994; Leea et al. 2008). A simplified representation of this process when this is performed by a machine is shown in  Fig. 50.4. The world model allows for the decision-making process to occur. The constraints reside on the machine representation of the world which is constructed by perception process that is based on the observations made from a series of vehicle onboard sensors, digital maps and the vehicle state (position, heading, speed, etc.)

The basis for building a world model originates from signals from the vehicle onboard sensors. To these signals different algorithms are applied in order to extract information



■ Fig. 50.4

World model in the vehicle navigation process

on features and on other entities sharing the road network (Sun et al. 2006). The models are built to take into account limits in the perception process, the uncertainty associated to the data used and the temporal properties. These can be represented in the form of occupancy grids using a probabilistic tessellated representation of spatial information. That is, a grid that corresponds to stochastic estimates of the occupancy states of the cells in a spatial lattice. For this purpose, probabilistic sensor models are used extensively. The computation of the grids can be done either at the lowest level, as in the case of the disparity space for stereo vision systems, information which is then transformed into a Cartesian space occupancy grid that forms part of the world model (Perrollaz et al. 2010). Within the robotics community there is a particular approach towards building a world model, this is known as the Simultaneous Localisation and Map Building (SLAM) problem, in which the robotic device as it moves builds a map of its environment while localising itself (Dissanayake et al. 2001). It is possible to combine this approach with the use of digital maps to facilitate the construction of the map and the localisation estimate (Lee et al. 2007).

A world model and the location of the subject vehicle in it is a basic requirement for autonomous vehicle navigation.

Motion. This function can be defined as a series of tasks (including path planning to reach the destination and for obstacle avoidance, and, and vehicle control) that enable the platform to move safely and efficiently. Determining the vehicle trajectory comprises two tasks. Local path planning that relates to the immediate motion of the vehicle for obstacle avoidance (Borenstein and Koren 1991). Global path planning indicates the path that vehicle is to follow from its current position to its destination using stored information on the road network and associated attributes. The results of the local planner are used to actuate the vehicle (Krogh and Thorpe 1986). The motion function in vehicle navigation answers to the question: “How can I do it?” It is important to note that determining the vehicle motion depends on the system capability to perceive as far as possible in order to anticipate situations that might represent risk. This is difficult if only onboard vehicle

sensors are used due to limits in their field of view and layout. Vehicle actuation is difficult due to safety constraints and cost. The vehicle heading and speed need to be controlled to create a path that avoids any obstacle perceived by the onboard sensors. Currently, the Electronic Stability Programmes (ESP) are the most used to control vehicle stability once sudden accelerations occur so as to avoid slippage. Longitudinal speed control included as part of Adaptive Cruise Control (ACC) that allows a vehicle to follow another at a safe distance is an example of computer controlled motion. Currently vehicle actuation is more and more under computer control, making the automation of vehicles more likely. Typical examples are the automated parking systems commercialised by several vehicle OEMs.

The Localisation, Mapping and Actuation functions have a high level of interaction. At the vehicle level, they conform to an interdependent complex system evolving in a highly unpredictable environment. Two observations can be asserted: A model of the world and its understanding is what determines the level of intelligence that can be embedded for decision making purposes. The construction of the model is limited by the size of the area which could be sensed, the limited field of view of the onboard vehicle sensors. Architecture for vehicle navigation that is centred on a representation of the world that the vehicle is to traverse is one of the basic requirements for automating vehicle navigation tasks. If the environment is known over large areas, it will be possible to anticipate what the vehicle could expect and plan accordingly.

A systematic representation that could respond to these requirements was proposed first by the 4-D/RCS architecture by J. Albus (Albus and Meystel 2001). It provides a layered representation of the world where each layer has different characteristics, in terms of size, accessing time, granularity of the information, etc. ● [Figure 50.5](#) shows a representation of this architecture and the manner in which the layers are distributed. Each layer represents different features such as road geometric primitives, object tracks, object groups, etc. The lowest layer level is the closest to the vehicle, a small area having a fine granularity. In general, all the data captured by the exteroceptive sensors is written into this area.

The decision making mechanism will scan this area at rapid intervals, the granularity should allow for the representation of vulnerable road users (pedestrians). The underlying structure for this layer is given by the geometry and attributes found in standard digital maps. Higher level layers will have larger zones of interest where objects will be identified and attributes associated to them. The refresh rates will be slower and the resolution coarse. Information from other vehicles or infrastructure will be in general written onto the upper layers so as to extend the situational awareness of the vehicle. The concept of structuring the world model as formulated by J. Albus has been applied in a landmark project on Cooperative vehicles safety applications, namely, the SafeSpot project as part of the Local Dynamic Map concept (SAFESPOT 2010). The later forms today part of a discussion on standards at the European Telecommunications Standard Institute (ETSI) where most of the standards for V2V and V2I applications are being developed. The Technical Committee ITS STF404 is addressing the standardisation of the Local Dynamic Map (LDM).

The different applications and technologies related to autonomous navigation reside on a representation of the human process that makes possible an understanding of the

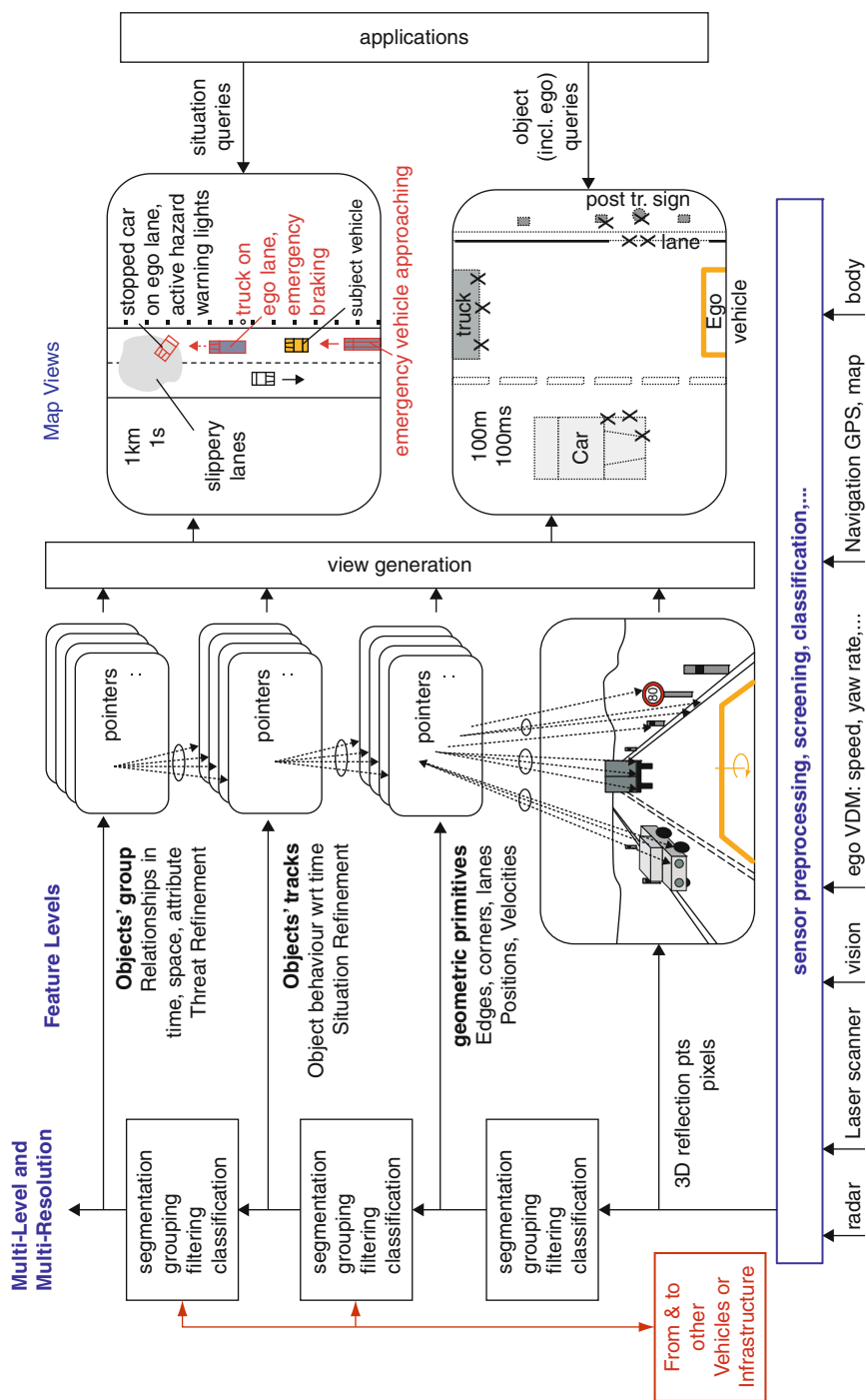


Fig. 50.5 4D-RCS world model and architecture for vehicle navigation (Albus and Meystel 2001)

vehicle situation with respect to its immediate environment. Anything that is not represented in this model will be ignored by the decision-making process and thus would lead to errors. Perception is a complex process that is limited by the physics of the sensors used, which leads to undefined areas, uncertainty in the measurements, and delays. One of the challenges in Intelligent Vehicle research is the construction of such a model and its interpretation in real-time under all types of driving conditions. The manner in which the first layer is represented is very important as actuation of the vehicle depends on decisions taken on the knowledge of the environment defined in this layer. *Occupancy grids* are used for this purpose as they encapsulate the multidimensional information stored on them in order to represent uncertainty (Jochem and Pomerleau 1995). Earlier work in this area considered that the road network was static; it was assumed that obstacles did not move. However in traffic conditions this is not the case. Today as vehicles have begun to move in cluttered environments, the dynamics of the obstacles and the limits of the perception systems are being incorporated (Fulgenzi et al. 2007). Concepts such as the probabilistic velocity obstacle (PVO) approach applied to a dynamic occupancy grid are being used in order to infer the likelihood of collision when uncertainty in position, shape and velocity of the obstacles, occlusions and limited sensor range constrain calculations.

3.1 Interaction

Different entities share the same road network; these include vulnerable road users, powered vehicles, powered two wheelers and bicycles. Their behaviour is determined by their interaction and the constraints imposed by traffic rules. That is, the interaction represents the spatio-temporal relationship between all entities, which has the underlying objectives to avoid collisions, reduce driver anxiety and ensure traffic flow.

As vehicles move in a road network, they interact with each other and other entities. This can be considered as a social phenomenon dependent on the emotional state, and physical conditions of the drivers, weather conditions and the layout of the road network. Thus interaction depends on the context. Interaction occurs with pedestrians, other powered vehicles, powered two wheelers and bicycles. Statistics have demonstrated that the interaction of pedestrians with passenger vehicles at intersections results in a high number of fatalities where pedestrian and driver demographic factors, and road geometry, traffic and environment conditions are closely related to conditions leading to accidents (Lee and Abdel-Aty 2005). Much work has been done in this area with results being incorporated into new passenger vehicles as in the case of Mobileye's Pedestrian Detection systems (Stein et al. 2010). However, the question resides not only pedestrian detection, but rather in the manner in which these interact with vehicles, how they move and react. Once pedestrians are detected, their future paths are difficult to predict as it is necessary to estimate the collision probability, so as to prevent any physical interaction with the vehicle (Gandhi and Trivedi 2008). This is a compound problem, when vehicles are close to pedestrians, it is likely that close gesture interaction occurs, for example a driver by

watching the eyes and direction of observation of a pedestrian can understand that the later is aware of its presence. This level of interaction is very difficult using current perception systems.

Driving a vehicle implies taking decisions continuously based on the current awareness of the vehicle situation and its likely evolution. Therefore the ability to infer the intentions of the actors in a scene from the available cues is essential. When estimating a driver's manoeuvre intention, it is necessary to account for interactions between vehicles. Indeed, the context in which a driver performs a certain action completely changes the interpretation that should be made of that action. For example, if a vehicle is changing lanes: information about other vehicles, their relative speeds, accelerations, etc. should facilitate the inference of the drivers' intentions when this is associated to the context, it should enable drivers to infer better the intentions of other vehicles and thus improve the inference of risk situations. Because of the high number of possible scenarios at road intersections, and the complexity of interactions between vehicles in these areas, driver intentions influence the decision making process. If a vehicle as it enters an intersection locates itself in the right hand lane and has activated its indicators, the computer controlling the observer vehicle will infer that it is highly likely that the driver has the intention to turn right; accordingly the behaviour will be different if there was no inference on the driver intention. Thus if manoeuvres of other vehicles can be predicted independently, it will be possible to estimate the collision risk and predict the manoeuvres that would avoid or reduce it. This is being applied to road intersection safety when using wireless communications technologies that enable the sharing of information amongst road actors (Lefevre et al. 2011).

Vehicles can be regarded as social entities, as such interaction is central to their behaviours, where compromises are part of the decision making process.

4 Classification of Technological Advances in Vehicle Technology

Vehicle navigation comprises the control of the mobile platform as it moves from its original position to its desired destination in a safe manner while traversing an infrastructure built for human driving. Given the state-of-the-art in sensing, computing and decision making, today a mobile platform would be able to cross most road networks, if it were the only user. The major difficulty resides in the sharing of the infrastructure with other entities like different powered mobile platforms, vulnerable road users, etc. The behaviour of these being unpredictable despite the existence of traffic rules and law enforcing mechanisms, driver errors occur, leading to a high number of road accidents. The major difficulty on deploying autonomous vehicles is on finding solutions that enable the sharing of the same workspace with other entities.

The *rationale* for studying vehicle navigation technologies applied to passenger vehicles is that *driver centric*, *network centric* and *vehicle centric* developments are all contributing to the development of autonomous vehicles.

4.1 Driver Centric

Today, the transport industry, universities and government R&D centres are developing Intelligent Vehicles from different perspectives. The car industry for example is deploying vehicle onboard technologies that facilitate the usage of a car by drivers and to improve safety. That is, the driver remains part of the control loop despite some tasks being delegated to sensor-based computer control systems. Two major issues define this strategy, cost and liability. The manner vehicles are perceived by the population has changed, these are not longer status symbols or trigger a passion up to the late 70s; today convenience, cost and usability are the main factors governing the purchase of a vehicle. While it has been demonstrated the potential of exteroceptive sensors such as laser scanners and infrared cameras for obstacle detection in standard traffic conditions, their use is limited to high-end vehicles only due to the costs involved. The reason is that prices are beyond what can be accepted in mid-price cars, where despite the limitations found in RADARS and video cameras, these are preferred. From this perspective advances in current mass-produced Intelligent Vehicles can be first defined as being *Driver Centric*.

Driver centric approaches enable the understanding of situations that autonomous vehicles will encounter when deployed in real traffic situations, it shows the techniques used on human controlled systems to increase safety are similar to those that will be tackled by autonomous vehicles. Further it allows for the transfer of technology from mass market vehicles to advanced experimental platforms and vice-versa which results in the engineering know-how of the vehicle OEMs. For example vehicles that include the automated parking assistance system have the infrastructure to perform sensor-based computer controllable manoeuvres, it can then be used to provide the interface, actuation and safety mechanisms to affect the longitudinal and lateral control of autonomous vehicles. Pioneer work on automated parking assistance (Paromtchik and Laugier 1996) has been performed in the mid nineties, and related products arrived on the market 6–7 years later.

4.2 Network Centric

The introduction of communications technologies onboard of passenger vehicles enables the sharing of information between vehicles (Vehicle to Vehicle, V2V) or between vehicles and the infrastructure (Vehicle to Infrastructure, V2I). This has led to different types of vehicles whose functionality resides on their integration onto a communications network that allows for V2V and V2I wireless links that are known as Cooperative Vehicles. In a functional manner these types of vehicles are regarded as *Network Centric*.



Network centric solutions are providing the means to share information among all actors in road networks. The network can then accumulate and analyse data prior to their broadcast to the networks. For autonomous vehicles this is a very important contribution as it means that the autonomous vehicles of the future *do not need to be stand-alone systems*. They are to be nodes that move in cooperation with other mobile nodes.

4.3 Vehicle Centric

A different approach consists on automating as much as possible the vehicle navigation functions. The vehicle comes under computer control and the role of the driver is reduced until it is no longer within the vehicle control loop; the vehicles are autonomous. The architectures replicate the function necessary to navigate a vehicle in an autonomous manner and thus all the system design is centred on the vehicle functions. These types of vehicle can be regarded as Vehicle Centric.

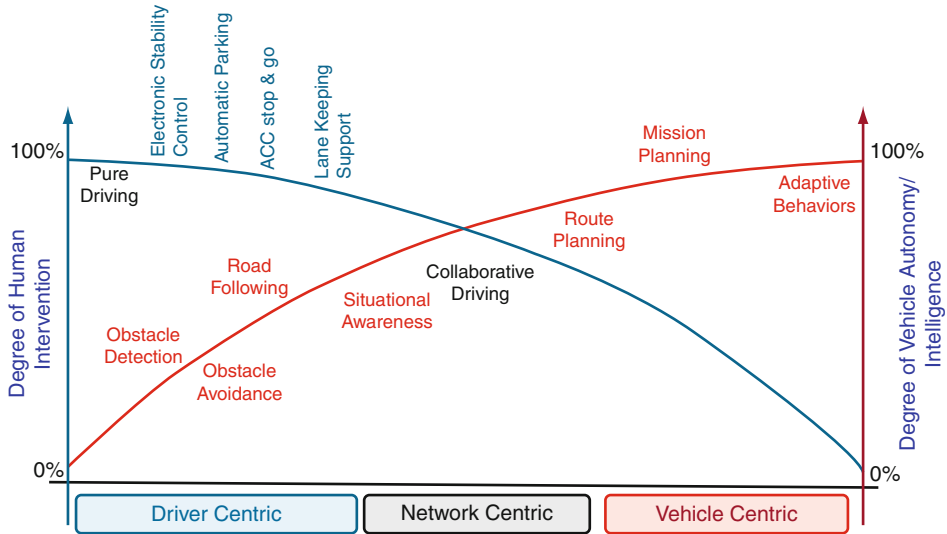
Vehicle Centric vehicles concern the realm of autonomous vehicles by considering the most salient experimental platforms developed so far and the associated technologies. This full panorama provides an overview of the technologies that are being developed for autonomous vehicles.

4.4 Evolution of Vehicle Architectures

From a vehicle controllability perspective, if there is 100% control by the driver, then, there will not be a vehicle navigation function under computer control. Most applications today are centred on informing the driver first and then letting the driver act on the vehicle, they are *driver centric*. Applications in which there is direct machine control are few, though these are being introduced gradually as in the case of Lane Keeping Support. Vehicle control is gradually being shifted away from the driver to computers. On a different perspective, computer controlled vehicles are gaining in autonomy from simple functions such as Obstacle detection and Obstacle avoidance to Situational Awareness and Mission Planning, with computers controlling ultimately the vehicle through various adaptive behaviours, *Vehicle Centric*. This shift is shown in  Fig. 50.6 where the transition between driver and computer (machine) control can be observed. Recent vehicle OEMs prototypes have demonstrated that the use of wireless links for safety-related driving functions are feasible as they enable the sharing of information and thus extend the driver awareness. These technologies should also be applicable to a *vehicle centric system* that is, communications networks lead to the development of intelligent spaces which facilitate the deployment of autonomous vehicles. The architectures issued from the use of communications systems are classified as *Network Centric*. A graphic representation of this transformation within the realm of Intelligent Vehicles is shown in  Fig. 50.6.

5 Driver Centric Technologies

In this category the overall control of the vehicle remains with the driver. Different functions are built to enhance the situational awareness of the driver and its safety either by improving the perception of the environment or maintaining the stability/controllability of the vehicle. All the perception functions are directed towards informing the driver about what occurs within its immediate environment. The emphasis is on the fusion of



■ Fig. 50.6

Evolution of vehicle architectures

data from proprioceptive and exteroceptive sensors that allow for the building of a model of the vehicle state and the spatio-temporal relationship with its immediate environment. This model is then used to infer situational information. Decisions are made all the time in every aspect of driving. As stated previously, accidents occur mainly due to driver error when the wrong decisions are taken.

The emphasis is on the acquisition and association of data to facilitate awareness and provide the most suitable means to enhance the driver situational awareness and hence facilitate decision making. In most cases this consists of perceiving what occurs within the immediate space in front of the vehicle or to detect its response. A typical example is the use of vision systems or laser scanners to detect pedestrians and either inform drivers or reduce the vehicle speed (Gerónimo et al. 2010; Gandhi and Trivedi 2006). This is a function that currently is being implemented as part of new generations of vehicles. The main difficulties reside on the perception systems due to the plethora of situation that might arise, the layout of sensors and their capabilities, with cost being a determining factor for deployment.

There are many applications aimed at assisting with the driving process. Several can be found on existing vehicles or are programmed into new products. Others have been demonstrated and are part of field operational tests. Today the preference is on *Informing* drivers. The rationale is linked to liability; that is, keeping the driver within the vehicle control loop means that the responsibility for the vehicle manoeuvres resides on the driver. Applications that include machine control of the vehicle include the Electronic Stability Programme (ESP), pre-crash braking (for mitigation purposes) and Adaptive

■ Table 50.2
Applications for driver centric systems

Route guidance (Navigation)	Longitudinal and lateral control	Enhanced perception
Navigation systems	Adaptive cruise control	Intelligent lighting control
Congestion avoidance	Automated driving functions (e.g., Parking)	Detection of aggressive drivers
Dynamic route guidance (e.g., congestion assistant)	Lane change/merge collision avoidance	Driver monitoring systems (e.g., attention, distraction, workload, drowsiness, drunk driving, etc.)
Electronic credentialing	Automated collision avoidance	Pedestrian detection
Electronic tolling	Intersection collision avoidance	Night vision
Black ice warning systems	Lane departure warning/keeping system	On-vehicle system monitoring
Highway geometry-based fuel burn optimization	Roadway departure	Around view monitor
Road/bridge condition monitoring	Platooning (E-Tow bar)	

Cruise Control. Their operation is based on the combined use of onboard sensors and on the incorporation of a priori information from navigation type digital maps. ➤ Table 50.2 presents a summary list of functions being implemented onboard current vehicles or demonstrators. The functions have been classified into three groups, namely Route Guidance (Navigation), Longitudinal and Lateral Control and Enhanced Perception. The first column comprises applications that rely on the use of digital road maps, the second addresses applications in which there is a degree of control on the vehicle motion and the third group includes applications that ameliorate the perception of the driver.

The applications listed in ➤ Table 50.2 implement active safety functions. By comparing the sensors, algorithms and strategies used in these applications, they are similar to those in many autonomous vehicles. *Driver Centric* vehicles are mainly implemented in passenger vehicles thus contrary to what occurs with *Vehicle Centric* vehicles that are mainly experimental platforms, cost and the layout of sensors are a major constraint. These applications must operate not only during demonstrations but under all type of traffic conditions. It is important to note that large investments are being made in R&D in this area, this should benefit the deployment of autonomous vehicles. The summary of applications shown in ➤ Table 50.2 lists the functions that autonomous vehicles should implement and the complexity of the challenge.

The overall structure of a *Driver Centric* architecture could be built around a model representation of the world surrounding the subject vehicle. Situation awareness information is generated by observing the model and inferring information that is transmitted to the driver. Mechanisms exist that allow for interaction with the driver, to adapt the entire system to the driver profile, to implement queries for information on the vehicle situation via the vehicle and world model, with traffic and other information incorporated into it via the connectivity mechanism. ➤ [Figure 50.7](#) shows the architecture of a vehicle centred on a driver.

Driver Centric architectures would lead to two ultimate functions that of operating a vehicle remotely or under the supervision of a central server as what occurs with the use of autonomous guided systems (AGVs) and the other will be of Indirect Driving.

Vehicle Tele-operation. It implies that sufficient information can be conveyed between a remotely located driver and the vehicle without compromising safety. This is an application domain preferred by the military for controlling land vehicles in hazard situations. ➤ [Figure 50.8](#) shows a typical purpose built remote control console used for controlling a vehicle at a distance (Ibanez-Guzman et al. 2004a). For this purpose means to observe the environment such as video cameras are used, to provide the sense of depth the return from laser scanners are deployed, this in conjunction with the video images are combined to generate depth coloured images. This type of system requires means to transport the vehicle controller as close to the vehicle and environment as possible. With the introduction of assisted steering, electromechanical braking and electrical traction, indirect observations of the environment like the around view monitor, etc.; the linkage between driver and vehicle is becoming indirect, that is there is no longer a direct physical contact, Vehicles are being driven indirectly, remotely, hence the interest by ergonomists on the use of technologies explored on the remote control of vehicles.

Indirect Vehicle Driving. This configuration has been explored by the defence organisations; the driver has contact with the outside environment and vehicle commands only by electro-optic means. That is cameras and other perception sensors are combined to provide a display of the driving environment. Instead of using the naked eye, the driver is assessing and interacting with the external world through the use of displays relaying live images from the camera system and other sensor data. It is thus possible to locate the driver anywhere within the vehicle, in the case of a military vehicle; the driver will be located in a protected area, representing a significant reduction in risk and costs. This overcomes the constraints associated to the need of locating driver on an advantageous observation point while remaining protected and being able to act on the vehicle commands (Ng et al. 2005). The availability of advanced perception systems in terms of scanning lasers and video cameras together with GPS and Inertial Measurements units facilitates the construction of detailed 3D views of the vehicle surroundings. Thus it is possible to present to the driver an extended panoramic view of the vehicle surroundings facilitating situational awareness and thus the driving of the vehicle in challenging situations. The Driver Awareness and Change Detection (DACD) system is a typical example in which a 3D 360° view of the environment around the vehicle is displayed to improve the driver situational awareness (DACD 2011).

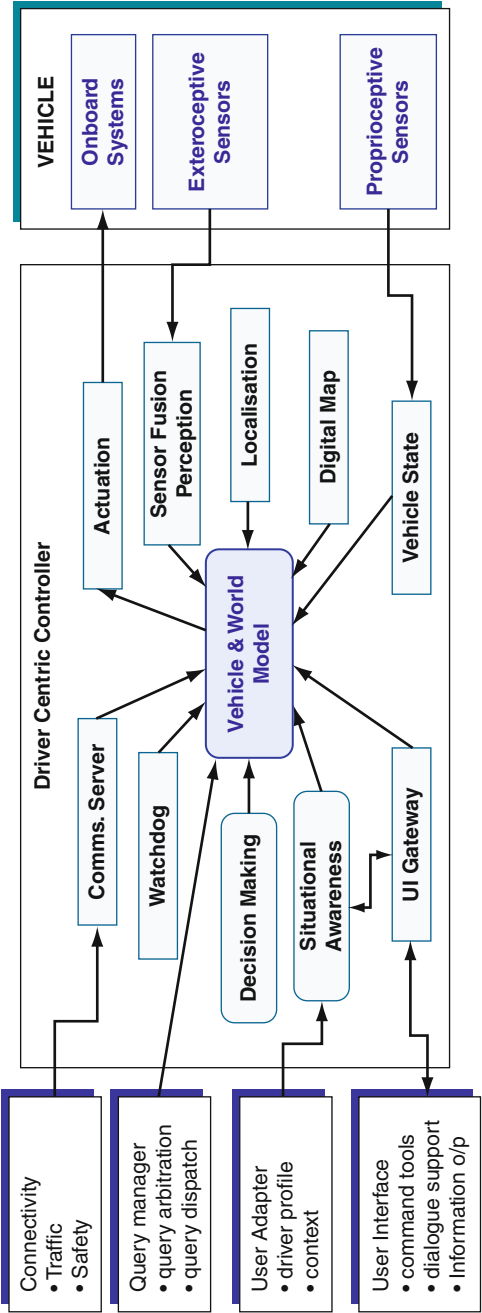
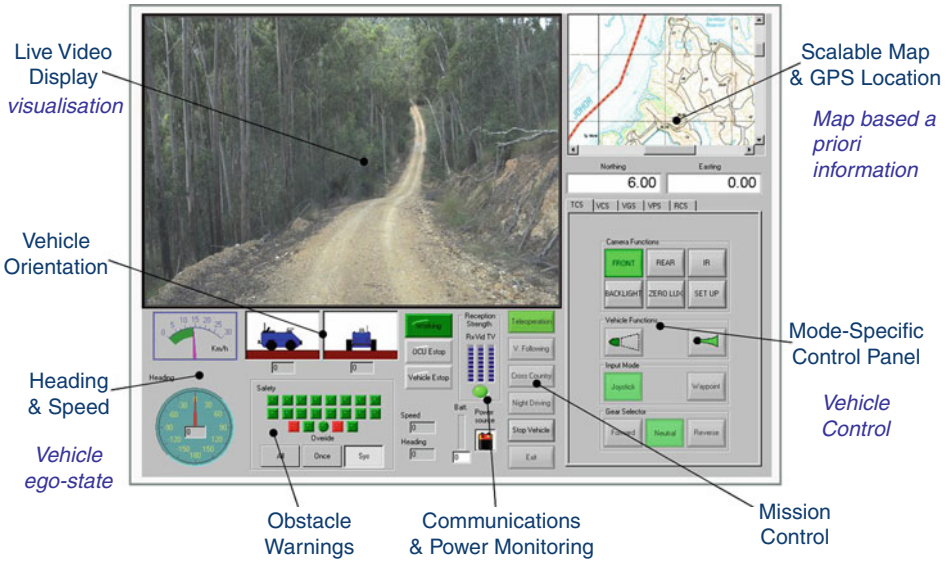


Fig. 50.7
Driver centric architecture (Ibanez-Guzman et al. 2004b)



■ Fig. 50.8

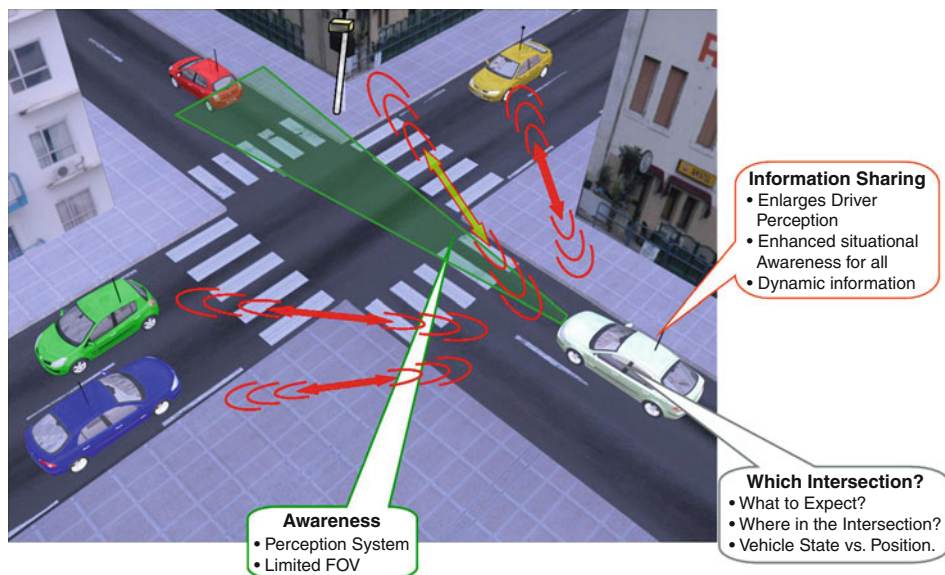
Console used for the remote control of a ground vehicle (Ibanez-Guzman et al. 2004)

6 Network Centric Technologies

Advances in computer and communications technologies are facilitating the establishment of vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) wireless communications links, both types of communications links are known as V2X. This connectivity is changing the way vehicles are designed (Mitchell et al. 2010). It enables the sharing and aggregation of information that results on an extension of the driver awareness horizon beyond what vehicle onboard sensors will perceive. From a safety perspective, it enables drivers or machines to react in advance with respect to the interaction with other mobile entities and to changes in environmental conditions (Lanson and Dauchez 2010; SAFESPOT 2010; Fukushima 2010; Martinez et al. 2010; NHTSA 2009).

The concept of *network centric vehicles* is based in three fundamental components: *Localisation*, a *Digital Map* and *Wireless Communications*. Information on the location and speeds of neighbouring vehicles and their projection onto a digital map that represents the road network and associated attributes, allows for the establishment of the spatio-temporal situation of other vehicles with respect to the subject vehicle (SV) and their interaction with the road network. This information is then used to infer whether or not risk situations arise and thus inform the driver or a computer controlled vehicle to react well in advance to what onboard vehicle sensors would allow due to the physical limits associated to them.

The problem originated by the limited field of view (FOV) of sensors and the contextual information that maps provide is shown in Fig. 50.8 applied to a road intersection. As the SV arrives at an intersection, if this has only a forward looking sensor in the

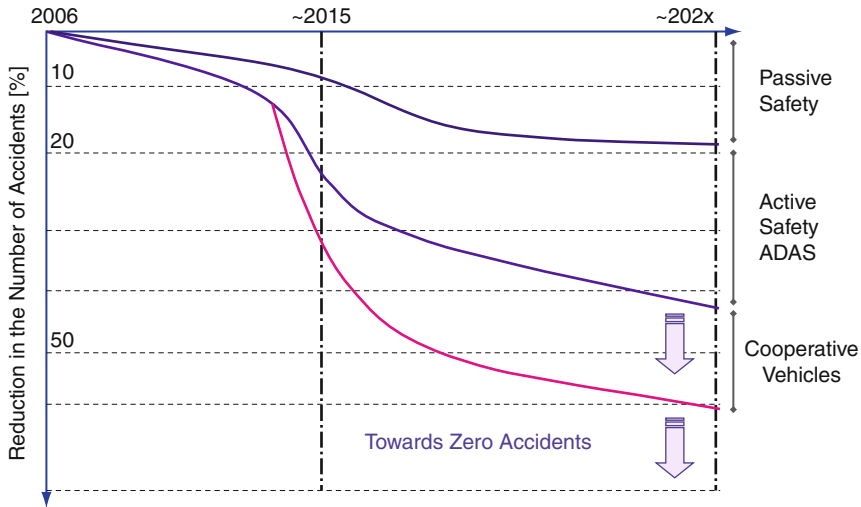


■ Fig. 50.9

Information flow at a road intersection from sensors, digital maps and wireless links

form of a radar, camera or scanning laser, the limited FOV of these sensors implies that incomplete information is available to the driver or computer controller. There is no awareness for example that the situation is more complicated as the SV will cross a road intersection or that the vehicle is arriving at a high speed and could not stop on time as the road traffic signal associated to its trajectory indicates. In this context, by incorporating information from the digital map it is possible to provide context to the data from the vehicle onboard sensor and improve awareness. ▶ [Figure 50.9](#) shows that either the infrastructure or other vehicles could transmit information to the SV. It is sufficient to transmit their state (position, speed plus other data) via V2X communications links thus the SV can have a model of the environment that includes the presence of other entities sharing the same road network.

At present there is much interest on the deployment of Cooperative Vehicles across the world by vehicle OEMs and governments. Statistics have shown that despite much progress in road safety, there is a slowing down on the reduction of accidents. Studies indicate that the next step in reducing safety is the use of cooperative vehicle safety applications (Biswas et al. 2006). ▶ [Figure 50.10](#) portrays the vision of the European Commission with regard to the use of V2X communications to reduce the number of road accidents. It states that to continue to reduce the number of accidents, it is not sufficient to deploy passive and active safety systems, that there is the need to start introducing safety V2X applications as these will reduce accidents that can not be done by conventional approaches. A part of this rationale is cost, vehicle onboard communications equipment, together with maps will be much cheaper than the use of complex sensors like radar and video cameras.

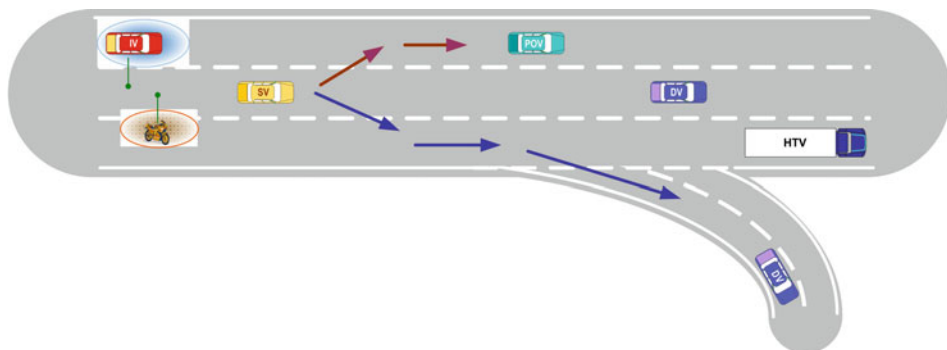


■ Fig. 50.10

Expected reduction in the number of accidents (Source EC, ICT for transport unit)

Interest on *network centric* intelligent vehicles in Europe is high; in 2010 the European Parliament has approved a directive to encourage the development and deployment of such systems within the European Community (European Parliament 2010). Numerous large R&D projects have been sponsored and the related technologies developed. Safespot (SAFESPOT 2010) and CVIS (2010) are examples of such projects. The latter addresses safety related applications and the former is oriented towards traffic flow and interaction with the infrastructure. The European Commission has allocated a frequency for V2X applications at 5.94 GHz. In Japan most vehicle OEMs are deploying similar systems, like for example the Sky project (Fukushima 2010). This is based on the V2I approach to increase safety at intersections and makes use of DSRC communications links that are widely used as part of the infrastructure in Japan. Other projects exist within this domain including Smartways promoted by the government to create a vehicle-road infrastructure that includes DSRC, navigation systems, road detectors, FM radio, etc. (Ministry Land Infrastructure 2011; Lanson and Dauchez 2010).

The principle of *network centric* vehicles relies on the time-stamped sharing of spatial information. This implies that the pose estimates of the vehicle together with the digital maps must have a high degree of precision for the successful deployment of cooperative vehicle applications. The importance of precise location estimates is shown in the scenario depicted in ● Fig. 50.11. In this case, if location information is transmitted to the SV from the intruder vehicle (IV) and the motorbike, their representation onboard the SV would indicate that overtaking or changing lanes can be dangerous due to the proximity of the neighbouring vehicles. However, if the position of either the motorbike or the IV is wrongly estimated, then, the cooperative vehicle application would provide the wrong information to the driver.



■ Fig. 50.11

Potential risks due to wrong location estimates

In *network centric* systems vehicles form ad hoc networks as they evolve over the road networks, if all the infrastructure nodes are interconnected distributed algorithms could be run in the network with capabilities to control the flow of all the mobile nodes within the network. Modern vehicles are equipped by electromechanical means to control their motion converting them into controllable nodes. By combining their onboard intelligence with that on the network, autonomous manoeuvres could be implemented (Srinivasan 2006). In this approach the context gathering and processing functions are distributed using sensor networks and wireless communications links to reduce the onboard intelligence and perception systems needed onboard vehicles to move safely either under partial driver control or computer control. That is a centralised system could gather information from multiple sources such as observation cameras, traffic lights, pedestrians (using smartphones) and moving vehicles. This information is then used to build an extended model of the environment on top of an accurate model of the road network. Information on all the entities then can be inferred with strategic decisions (e.g., the ordering of priorities at a road intersection) being downloaded as orders to the vehicles concerned, with operational decisions being made locally.

Autonomous vehicles when deployed would certainly use some of the technologies developed for cooperative vehicles as higher levels of automation are reached and wireless communications become a commodity. In economic terms it could be more attractive to replace onboard sensors with communications systems.

7 Vehicle Centric Technologies

When the control resides on the vehicle onboard processing systems, vehicles move in an autonomous manner thus developments are *vehicle centric*. All the perception functions operate in a manner that must be understood by a computer. That is, the vehicle onboard intelligence understands the vehicle situation and decides on the most appropriate manoeuvre as the vehicle moves towards its destination. The overall system architecture is designed for a class of autonomous systems incorporating perception, understanding,

decision making and actuation capabilities. While these endeavours were at first envisaged for defence or space applications, this is no longer the case. The correlation between these applications and those in the car industry was identified earlier on leading to a joint conference where researchers from the Department of Defence (DOD) and Department of Transportation (DOT) came together to share their endeavours in autonomous vehicles (IVVT 2011). The first were interested in the use of robotics technologies for defence applications while the second addressed improvements to the efficiency, safety and security of passenger vehicles.

The autonomous motion of ground vehicles can be traced back to the early days in robotics. One of the first examples is the *Stanford Cart* a card-table sized mobile robot that was equipped with a TV camera and transmitter. A computer was programmed to drive the cart through cluttered indoor and outdoor spaces, gaining its knowledge about the world entirely from images broadcast by the onboard TV system. The cart moved in 1 m spurts punctuated by 10–15 min pauses for image processing and route planning (Moravec 1980).

Understanding whether or not the vehicle can traverse the immediate environment, relies on perceiving the presence of obstacles along the desired path. The emphasis is on detecting obstacles and understanding the relation that exists between them and the SV. The availability of sensors to academia and industry has likely dictated the pace of developments, as autonomous mobile platforms moved away from laboratory environments, to well structured work spaces like warehouses, to off-road navigation, and to urban environments. These developments were associated to the availability of different sensors. From a vehicle manufacturer perspective, the sensors used in *vehicle centric* systems are very similar to those being introduced as part of driving assistance systems with cost being the major determining factor.

Robotics platforms using different perception systems can be found since the early 1980s, at first with an extensive exploitation of ultrasonic sensors and all limitations involved in terms of range and detection accuracy (Crowley 1985). Ultrasonic sensors are a standard feature in most current production passenger vehicles and are being used not only to assist during manoeuvres in tight enclosures but to guide autonomously vehicles during Parking Manoeuvres as for example the fully automatic parking system by INRIA (Paromtchik and Laugier 1996) or by Bosch GmbH. An early use of video cameras was made but this required purpose built computational hardware that precluded their intensive use, examples can be found in (Crowley 1985). Examples on the use of stereo vision to recover depth information for mobile robotics applications can be tracked back to the late 1980s (Ayache 1989). Video Cameras today are used as part of detection system onboard vehicles from rain detection to pedestrian detection.

Cost and performance concerns have been overcome and several vehicle OEMs are including them as part of their active safety systems. A notable example is monocular multipurpose camera by Mobileye. It is based on a vision-system on a chip, EyeQ2 which allows for the ability of multiple functions like lane departure warning, driver impairment monitoring, adaptive headlight control, pedestrian recognition, etc. Another important sensor for the development of autonomous *vehicle centric* systems has been Light Detection and Ranging (LIDAR) sensors known also under the name of Scanning Lasers or



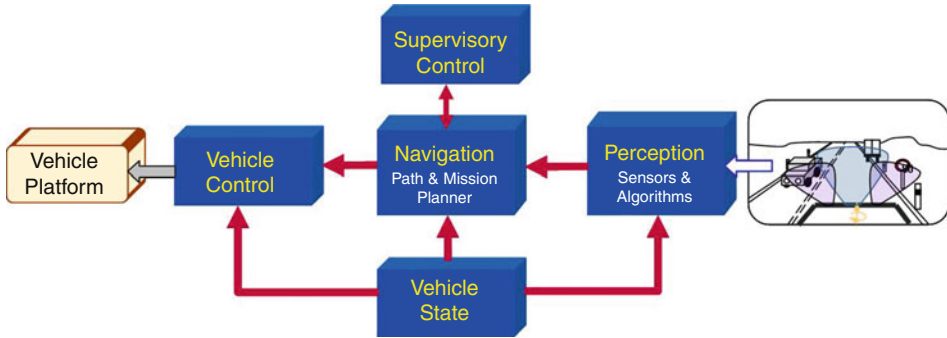
■ Fig. 50.12

(a) Single line, 180°. (b) 4 lasers/detectors 180°. (c) 64 lasers/detectors, 360°. Major scanning lasers deployed in autonomous ground vehicles

LADARS. Applications to navigation guidance can be traced to the early 1980s (Jarvis 1983). Over the years scanning lasers have been used as the main sensor for vehicle navigation, from single line scanning lasers to multiple lines and flash LADARS. Currently the benchmark unit is the Velodyne, a scanning laser which covers the vehicle surrounding environment with multiple vertical samples and detection range between 50 and 120 m. It has opened multiple possibilities to autonomous vehicle navigation as the perceived environment is more robust than what can be achieved using computer vision. ▶ Figure 50.12 illustrates the major scanning lasers deployed in autonomous vehicle applications. It should be remarked that for passenger vehicle applications, these sensors are seldom used, cost being the main constraint.

Experience has shown the limitations of infrared and optical systems in dust or fog conditions, thus another active sensing technology is also used, namely millimetre wave radar in the 24 and 77 GHz bands for distance measurement. The application of radar to autonomous vehicles can be traced to the early 1990s (Landge and Detlefsen 1991). These sensors have been used beyond the obstacle detection, to locate vehicles in unknown environments as the Radar based simultaneous localisation and map building (SLAM) (Mulane et al. 2007). However their use in passenger vehicle applications has been limited to Adaptive Cruise Control (ACC) systems mainly due to costs and the manner in which they can be integrated onboard these vehicles.

A simplified functional architecture associated to *Vehicle Centric* systems is shown in ▶ Fig. 50.13. Essentially there are four interdependent systems: The *perception system* comprises sensors used to capture in general radiant energy in the scene (exteroceptive sensors), complex algorithms are used to detect, locate and recognise a potential feature of interest. The *navigation system* includes a series of algorithms used for situation understanding, decision making and to generate the path that the vehicle should follow. This is based on a model built from a perception of the environment and on information on the vehicle state and a priori digital map information as represented in ▶ Figs. 50.4 and ▶ 50.5. The *control system* receives commands from the navigation system in terms of



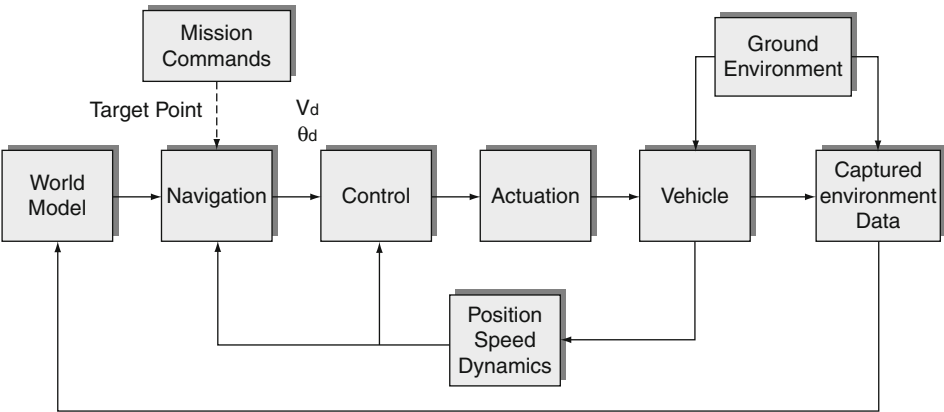
■ Fig. 50.13
Functional diagram of a vehicle centric architecture

heading and speed; its function is to ensure that the vehicle follows the desired trajectory. It includes means to act on the vehicle (directly via existing actuators or directly on the controllers of the traction systems). In addition, to maintain coherence of the underlying systems and ensure that these navigate safely or fail-safe, a *supervisory control system* is included to perform this function. It consists of monitoring mechanisms and processes that are triggered in case of performance degradation or malfunction of any of the systems. Most functions depend on the status of the vehicle itself, that is, its position, speed, acceleration, yaw rate, etc. This information is used in the planning phase, to provide a better understanding of the vehicle situation within the machine-understandable world model. This is known as the estimation of the *vehicle state*, which comprises means to estimate the vehicle dynamics plus its position with respect to a global reference frame as well as in a local manner.

Different instances of *vehicle centric architectures* exist. These have evolved over time, though the basic functions remain the same. It is fundamentally a control system made of several loops, the inner one as related to the actuators acting on the commands of the vehicle like steering mechanism or propulsion systems, the next outer loop will be the control system that ensures that vehicle response is as desired in the navigation system and the outer control loop which is controlled by the navigation system associated to the generation of trajectories in accordance with the perceived environment and vehicle state. ♦ *Figure 50.14*, shows a simple control system representing the functions needed in a *vehicle centric* family of autonomous vehicles. If the vehicle has wireless links, it can communicate with other vehicles and the infrastructure, the shared information will be written into the world model and decisions made locally as part of the path planner. Shared information enriches the world model hence the interest on networked vehicles.

8 State-of-the-Art in Fully Autonomous Driving Research

Fully autonomous driving has long been envisioned in science fiction (such as the film *Minority Report*) and in the robotics field. The PATH project (Horowitz and Varaiya 2000)

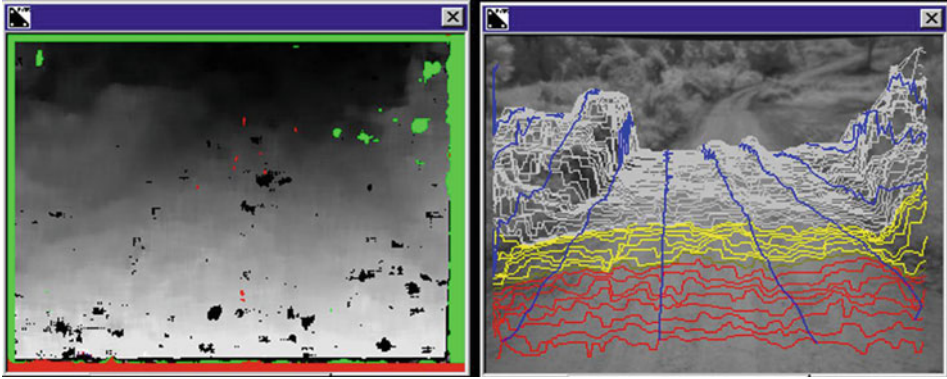


■ Fig. 50.14
A control system level representation of a vehicle centric vehicle

as well as Navlab and “Hands-free across America” (Jochem and Pomerleau 1995) were among the early projects in the area of fully autonomous driving. Fully autonomous driving is attractive because of the enormous potential for increased safety, productivity, and fuel efficiency. Safety would be enhanced because of the reduction of driver-caused accidents. Productivity would be increased because humans in the car could work while commuting, and because vehicles could safely travel much faster. Fuel efficiency would benefit because efficient routes could be planned on-the-fly, and because fully automated vehicles could drive very closely to one another, reducing the effect of air resistance and greatly increasing highway throughput. Fully autonomous driving requires many components to function, such as perception, navigation, control, etc. Various sections of this handbook are focusing on current research in each of these areas. Here the focus is specifically on projects which have advanced the state-of-the-art for fully autonomous driving.

8.1 Early US Unmanned Vehicle Projects

An important milestone and benchmark were the first NAVLAB vehicles at Carnegie Mellon University, platforms that were used to develop, integrate and test advanced technologies for autonomous vehicle guidance. Multiples concepts and experiments were developed and efforts can be considered a landmark in the domain (Thorpe 1990). NAVLAB vehicles could be driven in road networks; they were used to introduce the concept of occupancy grids, the development of path planning algorithms, etc. (Elfes 1989; Stentz 1994). As part of defence applications very close related technologies were developed around the same period. A landmark programme was Demo III Experimental Unmanned Vehicle (XUV) program which centred on technological developments, modelling, simulation and experimentation as well as the integration of the resulting



■ Fig. 50.15
Disparity map generated from a stereo vision system, interpretation of the data to determine the likely free-obstacle paths

system with the users. Vehicles were moved at speeds of up to 32 km/h during daylight and 16 km/h at night by the end of 2001 (Schoemaker and Borsntein 1998). Within this program extended experiments on the use of stereovision, infrared and colour cameras, and scanning lasers were made. Early work showed the potential and difficulties encountered on the use of video cameras to detect range, mainly using stereo vision. ► *Figure 50.15* shows the disparity map generated from a stereo camera pair which then it had to be processed to understand the data and infer information on what is the ground and potential obstacles. At that time the centre of interest was on generating an occupancy grid representing *traversable areas* and *perceived obstacles*. The difficulties on using vision to measuring depth were also highlighted and the advantages of using active sensors like scanning lasers.


8.2 Early EU Projects on Intelligent Vehicles

The two final demonstrator vehicles in the European project Prometheus (PROgramMme for a European Traffic of Highest Efficiency and Unprecedented Safety, 1987–1995): VITA_2 and VaMP may well be considered as the first two road vehicles of a new type capable of understanding their environment and of reacting appropriately to different situations in an autonomous manner (Dickmanns 1997). Prometheus was, at the time, the largest R&D project in the field of driverless cars. It was funded by the European Commission and defined the state-of-the-art in autonomous vehicles. The project is likely at the origin of different concepts that were advanced further in various platforms. The extensive interest in Europe on active safety systems and their widespread development probably has its roots in Prometheus. In the early 2000s, interest on unmanned ground vehicles was intensive in several countries. A classic example is the German experimental robotics program PRIMUS (PRogram for Intelligent Mobile Unmanned Systems) that

addressed unmanned manoeuvres in unknown open terrain. The emphasis was on the development of algorithms for a high degree of autonomous navigation skills with commercial off-the-shelf devices (COTS) integrated onto military vehicles. The guidance of such vehicle relied very much on the use of a 3-D Ladar for obstacle detection and a colour camera for road detection (Schaub et al. 2004). A similar approach was taken for a project in South-East Asia in which an armoured tracked vehicle was converted onto an unmanned ground vehicle for operation in tropical environments using low-cost COTS systems (Ibañez-Guzmán et al. 2004b). These platforms demonstrated the potential and feasibility of deploying unmanned vehicles in challenging environments using existing technologies.

During the last 5 years, several approaches have been developed for dealing with two main related issues: dealing with robustness and efficiency problems for vehicle perception (Coue et al. 2006), and dealing with dynamic and uncertain environments (Laugier et al. 2007).

8.3 US DARPA Challenges on Autonomous Driving

Recent work in fully autonomous driving was focused by the Defense Research Advance Projects Agency (DARPA) – sponsored Grand Challenges. These three events were part of the impetus for significant improvement in this field. The majority of the details that follow are found in (Buehler 2009). The first Grand Challenge took place March 13th, 2004. One hundred and seven teams registered for the event, and from this 15 teams took part in the final race. The race was a 142 mile off-road course with a requirement to complete the course in 10 h or less. None of the participants succeeded – in fact, no competitor completed more than 5% of the course. The second Grand Challenge took place a year and a half later, on October 8th, 2005. The challenge was very similar to the first – this time 195 teams registered and 23 took part in the race. Five of those teams completed the course, with Stanford's entry "Stanley" winning the race with a finishing time under 7 h, earning the 2 million dollar prize. Details of the equipment used in "Stanley" are shown in  Fig. 50.16, namely the GPS receivers and a series of single-line Laser scanners.

The 3rd DARPA Grand Challenge, also called the Urban Challenge, significantly increased the difficulty of the event, two major changes were introduced: First, the vehicles were in an urban setting which required complying with California traffic rules. Secondly, in the urban setting, the vehicles would have many more interactions with other agents, including other autonomous vehicles, parked vehicles, and human-driven vehicles. The event required teams to prequalify by having their vehicles tested and performing a series of tests. Vehicles had to prove they could follow basic traffic laws and complete a series of maneuvers, including parking. For a full description of the qualification procedures, see [DARPA]. Of the 89 teams that were registered, only 11 qualified for the final event. Six of these vehicles finished the challenge, with "Boss," Carnegie Mellon University's entry, winning the challenge. A very good technical overview of the various competitors and their technical approaches is found in (Buehler 2009). This book includes a paper



■ Fig. 50.16

Equipment used on “Stanley”, at the 2nd DARPA grand challenge

examining all of the robot-robot interactions from the point of view of the robots, and investigates the cause of the accidents at the event. It is to note that five of the six vehicles that completed the challenge made use of an off-the-shelf three-dimensional laser scanner from Velodyne, a technology which was not available when the first Grand Challenge took place.

8.4 EU Cybercars Projects

By no means was the Urban Challenge the only activity in fully autonomous driving over the last decade. While a full review of all work is not possible here, a few other efforts will be highlighted. Cybercars (Parent and De La Fortelle 2005) is an ongoing European effort to develop fleets of fully automated vehicles moving in town centres. Deployments to date have been in areas without human-driven vehicles, but in areas with pedestrian traffic at very slow speeds.

It should be noted that in some limited contexts, fully autonomous driving is already a reality. In addition to the limited deployments of Cybercars, fully automated test drivers are in use at the Transportation Research Center (Mikesell et al. 2008) and at Mercedes Benz (2010). This allows for more repeatable, safer testing of cars through standardized safety and performance tests. This field has been ongoing, with a summary of earlier efforts on automated testing well described in (Schmidt et al. 2000; Shoval et al. 1998). Clearly this demonstrates the ability for accurate control of the vehicle, but these systems require little to no perception as they evolve in highly-controlled environment with little

or no obstacles. Human intervention is required if there are any changes to the environment.

There has also been extensive work in special cases of driving. For instance, a variety of papers have dealt with the problem of autonomously parking a vehicle for over a decade (Paromtchik and Laugier 1996; Gorinevsky et al. 1994). It is interesting to note that this is one technology that has already transferred from research into some production vehicles, such as the Toyota Prius (Suhr et al. 2010). A much more dynamic autonomous parking demonstration was described in (Kolter et al. 2010). Other specialty areas of driving include driving in parking structures (Kummerle et al. 2009; Pradalier et al. 2005), driving off-road (Stentz et al. 2007), etc.

8.5 Lessons Learned from DARPA Challenges

Of particular interest in Buehler (2009) are the “lessons learned” from the grand challenge. A few of those will be mentioned here, as they given an overview of areas for additional work.

- Off-the-shelf sensors have made significant advancement (such as the Velodyne), but currently are insufficient for truly autonomous driving. In addition, the cost, packaging, and reliability in electromagnetically-noisy areas are not currently practical for production vehicles.
- Better representations of the environment will result in better performance. This includes more accurate representations of distances and speeds, better classification of objects, etc.
- Validation is not currently possible when interacting with the physical world. All of the teams used a variety of ad-hoc rules and scenarios, which worked with varying degrees of success. But because of the complexity of interacting with the physical, dynamic environment, it is currently impossible to prove completeness or correctness. This has major implications for liability.
- Human drivers adapted very well to the autonomous vehicles.
- Keeping a human-in-the-loop to respond to unexpected scenarios and having the vehicles operate semi-autonomously greatly reduces the complexity of the situation.
- During the Urban Challenge, there were no traffic lights, pedestrians, bicycles, and humans often had to pause vehicles to avoid potential accidents or other problems. None of these simplifications are realistic for fully autonomous driving in an urban setting.
- Time-to-collision as the primary means of detecting risk has some drawbacks, particular in cases where a vehicle is temporarily stationary (i.e., at an intersection). This can cause very high time-to-collision values while in fact the risk can be high. This type of error was involved in causing a collision and several near-collisions during the event. Current work on this issue is focusing on probabilistic collision risk assessment (Laugier et al. 2009; Lefevre et al. 2011).



■ Fig. 50.17
The google autonomous car

By no means has development in this area stopped since the Urban Challenge. Very recently, the Google Cars project has become public knowledge (Markoff 2010; Thrun 2011). The Google Cars are Toyota Prius passenger vehicles equipped with a 360° multilayer laser scanner and vision systems, plus actuation mechanisms to control their heading and speed. ● Figure 50.17 shows one of the cars with its standard equipment. This project has taken some of the technology developed for the Urban Challenge, and has been deployed in normal traffic on California roads. As of March 2011, over 140,000 miles have been driven using seven vehicles, with limited human intervention. While there is no short-term plan to commercialize this technology, this project is continuing to bring fully autonomous driving closer to an everyday reality.

9 Discussion, Conclusions and Structure of the Chapters

This chapter has reviewed the current state-of-the-art in Intelligent Vehicles and associated technologies. A panorama was presented on advancements in this domain; this was classified according to the emphasis given to the overall structure of the solutions proposed. It has been largely shown that the technologies for the deployment of autonomous vehicles in real traffic situations potentially exist in terms of sensors and algorithms. Developments have been largely led by the availability of depth range sensors used to observe the immediate environment and the algorithms used to extract information on the spatio-temporal relation of the vehicles and their environment. Detecting and identifying obstacles is a well understood problem and the necessary sensors and algorithms

exist even if reliability and efficiency issues have still to be improved. The same applies to the estimation of the vehicle position, orientation, speed as well as the use of a priori information in the form of digital maps. That is the estimation of vehicle state and its context.

Demonstrators have shown that the use of wireless communications technologies to share information between the various actors and the infrastructure should facilitate the deployment of such systems as the perceived area could be beyond what standard onboard sensors could attain.

The complexity involved on integrating multiple complex systems and their interdependency, means that autonomous vehicles are “system of systems” and thus their development needs to be addressed at that level if they are to be deployed successfully. At present, across the automotive industry there is a widespread application of Systems Engineering to the entire process due to the complexity encountered in current vehicles (Durrant-Whyte 2001).

The major challenge on Intelligent Vehicles resides on understanding the spatio-temporal relationship of the vehicle and its environment and on predicting the likely behaviour of the entities sharing the same work space as the vehicle. System integrity is another issue, if vehicles are to move autonomously close to humans more likely as *social networked platforms* rather than standalone systems. Another issue is on the manner in which embedded uncertainty could be reduced in the estimations used as part of these vehicles.

Today vehicles are becoming more and more self-reliant; mission-critical driving decisions are being done under computer control. Therefore, public policies, legislation, and technical standards are needed to prepare courts and the public for the new realities of traffic with autonomous or semi-autonomous vehicles. This is an important issue as it is now possible to control longitudinally and laterally the motion of the SV without driver intervention. A think-tank, the Centre for Automotive Research at Stanford (CARS), Stanford University has started to look into this issue from a legal perspective in cooperation with vehicle OEMs. The question resides on situations where drivers are left out from the control loop (Beiker and Calo 2010).

To conclude, efforts on active driving assistance systems that started as far back as the 1960s are converging to what are now known as autonomous vehicles. While there is still progress to be made in terms of cost, robustness, and legislation, recent results have shown of the technical feasibility and potential for improving safety, efficiency and to satisfy the mobility needs in the twenty-first century. Three complementary approaches have been discussed: those providing solutions to assist drivers, those relying on the use of wireless networks, and those moving towards autonomous vehicles. The combination of these three approaches should change mobility in the next years as vehicles become interconnected and the needs for sustainable means of transport become a necessity. In the research realm system integrity remains a major issue, as vehicle need to operate beyond simple demonstrators, their operation needs to be in all weather and traffic conditions and potentially becomes nodes of an interconnected world with strong interdependencies.

References

- Albus JS, Meystel AM (2001) Engineering of mind: an introduction to the science of intelligent systems. Wiley, New York
- Ayache N (1989) Artificial vision for mobile robots. MIT Press, Cambridge, MA
- Beiker SA, Calo MR (2010) Legal aspects of autonomous driving, the need for a legal infrastructure that permits autonomous driving in public to maximize safety and consumer benefit, Center for Automotive Research at Stanford (CARS), Stanford University, Stanford, Oct 2010
- Biswas B, Tatchikou R, Dion F (2006) Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety, IEEE Communications Magazine, Jan 2006
- Borestein J, Koren Y (1991) The vector field histogram-fast obstacle avoidance for mobile robots. IEEE J Robot Autom 7(3):278–288
- Buehler M, Iagnemma K, Singh S (eds) (2009) The DARPA urban challenge: autonomous vehicles in city traffic, Number 56 in springer tracts in advanced robotics. Springer, Berlin
- Coué C, Pradalier C, Laugier C, Fraichard T, Bessière P (2006) Bayesian occupancy filtering for multitarget tracking: an automotive application. Int J Robot Res 25(1):19–30
- Crowley JL, (1985) Dynamic world modeling for an intelligent mobile robot using a rotating ultrasonic ranging device. In: Proceedings of the IEEE international conference on robotics and automation, Pittsburgh, p 128
- CVIS (2010) Cooperative vehicles infrastructure systems. <http://www.cvisproject.org/>. Accessed 8 Mar 2010
- Dickmanns E (1997) Vehicles capable of dynamic vision. In: Proceedings of international joint conference on artificial intelligence, Neubiberg, pp 1577–1592
- Dissanayake G, Newman P, Durrant-Whyte H, Clark S, Csobza M (2001) A solution to the simultaneous localisation and map building (SLAM) problem. IEEE T Robot Autom 17(3):229–241
- Durrant-Whyte H (2001) A critical review of the state-of-the-art in autonomous land vehicles, systems and technologies, Sandia report, SAND2001-3685, Sandia National Laboratories, Albuquerque
- Elfes A (1989) Using occupancy grids for mobile robot perception and navigation. Computer 22(6):46–57
- European Council for Automotive R&D, EURCAR, (2010) R&D directions. www.Eucar.be. Accessed 16 Dec 2010
- Fouque C, Bonnifait P (2009) On the use of 2D navigable maps for enhancing ground vehicle localization. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, IROS 2009, St. Louis, pp 1885–1890
- Fukushima M (2010) Realization of NISSAN cooperative ITS systems – for traffic safety and traffic congestion, 13th international IEEE annual conference on intelligent transportation systems, Porto, Sept 2010
- Fulgenzi C, Spalanzani A, Laugier C (2007) Dynamic obstacle avoidance in uncertain environment combining PVOs and occupancy grid. In: Proceedings of the IEEE international conference on robotics and automation, ICRA 2007, Roma, pp 1610–1616, Apr 2007
- Gandhi T, Trivedi MM (2006) Pedestrian collision avoidance systems: a survey of computer vision based recent studies. In: Proceedings of the IEEE international transportation systems conference, Toronto, Sept 2006
- Gandhi T, Trivedi M (2008) Computer vision and machine learning for enhancing pedestrian safety. In: Computational intelligence in automotive applications, vol 132, Studies in computational intelligence. Springer, Berlin, pp 79–102
- Gerónimo D et al (2010) Survey of pedestrian detection for advanced driver assistance systems. IEEE T Pattern Anal 32(7):1239–1258
- Gorinevsky D, Kapitanovsky A, Goldenberg A (1994) AUTOPASS: automated parking support system, SAE technical paper 94100, Warrendale, 1994
- Guivant J, Nebot E (2001) Optimization of the simultaneous localization and map-building algorithm for real-time implementation. IEEE T Robot Autom 17(3):242–257
- Horowitz R, Varaiya P (2000) Control design of an automated highway system. Proc IEEE 88(7):913–925
- Ibañez-Guzmán J et al. (2004) Unmanned ground vehicles for natural environments. In: Proceedings of the 24th ASC2004, Army science conference, Orlando, Nov 2004

- Ibañez-Guzmán J, Xu J, Malcolm A, Gong Z, Chan CW, Tay A (2004) Autonomous armoured logistics carrier for natural environments. In: Proceedings of the IEEE international conference on intelligent robots and systems (IROS), Sendai
- Ibañez-Guzmán J, Lefevre S, Rodhaim S (2010) Cooperative vehicles to facilitate the crossing of intersections by emergency service vehicles. In: Proceedings of the international conference: V.I. S.I.O.N., Montigny le Bretonneux, 6–7 Oct 2010
- Intelligent Vehicle Technology Transfer (IVTT) Program (2011), <http://www.intelligent-vehicle.com/>. Accessed 26 Mar 2011
- International Road Assessment Programme, IRAP (2007) The true cost of road crashes: valuing life and the cost of a serious injury
- Jarvis R (1983) A laser time-of-flight range scanner for robotic vision. *IEEE T Pattern Anal* PAMI-5(5):505–512
- Jochem T, Pomerleau D (1995) No hands across America official press release. Carnegie Mellon University, Pittsburgh
- Kolter JZ, Plagemann C, Jackson DT, Ng AY, Thrun S (2010) A probabilistic approach to mixed open-loop and closed-loop control, with application to extreme autonomous driving. In: Proceedings of the international conference on robotics and automation (ICRA), Anchorage
- Krogh B, Thorpe C (1986) Integrated path planning and dynamic steering control for autonomous vehicles. In: Proceedings of the IEEE international conference on robotics and automation, Orlando, pp 1664–1669, May 1986
- Kummerle R, Hahnel D, Dolgov D, Thrun S, Burgard W (2009) Autonomous driving in a multi-level parking structure. In: Proceedings of the international conference on robotics and automation (ICRA), Kobe
- Lange M, Detlefsen J (1991) 94 GHz three-dimensional imaging radar sensor for autonomous vehicles. *IEEE T Microw Theory* 39(5):819–827
- Langer D, Rosenblatt JK, Hebert M (1994) A behavior-based system for off-road navigation. *IEEE T Robot Autom* 10(6):776–783
- Lanson F, Dauchez P (2010) Les Systèmes de Transport Intelligentes au Japon, Ambassade de France au Japon, Nov 2010
- Laugier C, Petti S, Vasquez D, Yguel M, Fraichard Th, Aycard O (2007) Steps towards safe navigation in open and dynamic environments. In: Laugier C, Chatila R (eds) *Autonomous navigation in dynamic environments*, Springer tracts in advanced robotics (STAR). Springer, Berlin
- Laugier C et al. (2009) Vehicle or traffic control method and system. Patent no. 09169060.2 – 1264, Aug 2009. Patent registered with Toyota Europe
- Le Marchand O et al. (2008) Performance evaluation of fault detection algorithms as applied to automotive localization. In: Proceedings of the European navigation conference, global navigation satellite systems ENC GNSS, Toulouse
- Le Marchand O, Bonnifait P, Ibañez-Guzmán J, Bétaile D, Peyret F (2009) Characterization of GPS multipath for passenger vehicles across urban environments. In: Proceedings of the European navigation conference, global navigation satellite systems ENC GNSS, Naples
- Lee C, Abdel-Aty M (2005) Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. *Accid Anal Prev* 37:775–786, Elsevier
- Lee KW, Wijesomaa S, Ibañez-Guzmán J (2007) A constrained SLAM approach to robust and accurate localisation of autonomous ground vehicles. *Robot Auton Syst* 55(7):527–540
- Leea J, Forlizzia J, Hudsona SE (2008) Iterative design of MOVE: a situationally appropriate vehicle navigation system. *Int J Hum-Comput St* 66(3):198–215
- Lefèvre S, Laugier C, Ibañez-Guzmán J, Bessière P (2011) Modelling dynamic scenes at unsignalised road intersections, INRIA research report 7604, INRIA Grenoble Rhône-Alpes
- Levinson J, Thrun S (2010) Robust vehicle localization in urban environments using probabilistic maps. In: Proceedings of the IEEE conference on robotics and automation (ICRA), Anchorage
- Markoff J (2010) Google cars drive themselves, in traffic. *New York Times*, 9 Oct 2010
- Martinez FJ, Toh CK, Cano JC, Calafate CT, Manzoni P (2010) Emergency services in future intelligent transportation systems based on vehicular communications networks, IEEE intelligent transportation systems magazine, 6, Summer
- Mercedes-Benz (2010) Mercedes-Benz automated driving. <http://www.popularmechanics.com/cars/news/industry/mercedes-puts-car-on-autopilot>. Accessed 10 July 2010
- Mikesell D, Sidhu A, Heydinger G, Bixel R, Guenther D (2008) Portable automated driver for road

- vehicle dynamics testing, ASME dynamic systems and control conference, Ann Arbor
- Ministry of Land, Infrastructure and Transport (2011) ITS introduction guide - shift from legacy systems to smartway. <http://www.hido.or.jp/itsos/images/ITSGuide-2.pdf>. Accessed 20 Mar 2011
- Mitchell WJ, Borroni-Bird CE, Burns LD (2010) Reinventing the automobile: personal urban mobility for the 21st century. MIT Press, Cambridge, MA
- Moravec H (1980) Obstacle avoidance and navigation in the real world by a seeing robot rover, PhD dissertation, Stanford University
- Mullane J et al (2007) Including probabilistic target detection attributes into map representations. *Int J Robot Auton Syst* 55(1):72–85
- National Highway Traffic Safety Administration, NHTSA (2008) National motor vehicle crash causation survey: report to congress. DOT HS 811 05, July 2008
- National Highway Traffic Safety Administration, NHTSA (2009), Vehicle safety communications—Applications VSC-A, 2009. <http://www.nhtsa.dot.gov/staticfiles/DOT/NHTSA/NRD/Multimedia/PDFs/Crash%20Avoidance/2009/811073.pdf>. Accessed 17 Nov 2009
- Ng KB, Tey HC, Chan CW (2005) Autonomous unmanned ground vehicle and indirect driving, in DSTA horizons. DSTA, Singapore
- Parent M, De La Fortelle A (2005) Cybercars: past, present and future of the technology. In *Proceedings of the ITS World Congress*, San Francisco
- Paromtchik IE, Laugier C (1996) Autonomous parallel parking of a nonholonomic vehicle. In: *Proceedings of the IEEE international symposium on intelligent vehicles*, Tokyo, pp 13–18
- Perrollaz M, Yoder J-D, Spalanzani A, Laugier C (2010) Using the disparity space to compute occupancy grids from stereo-vision. In: *Proceedings of 2010 IEEE/RSJ international conference on intelligent robots and systems, IROS 2010*, Taipei, pp 2721–2726
- Pradalier C, Hermosillo J, Koike C, Brailion C, Bessière P, Laugier C (2005) The CyCab: a car-like robot navigating autonomously and safely among pedestrians. *Robot Auton Syst J* 50(1):51–68
- Quddus MA, Washington YO, Noland RB (2007) Current map-matching algorithms for transport applications: state-of-the art and future research directions. *Transport Res C* 15:312–328
- SAFESPOT (2010) Cooperative systems for road safety. <http://www.safespot-eu.org/>. Accessed 3 Mar 2010
- Schaub G, Pfaendner A, Schaefer C (2004) PRIMUS: autonomous navigation in open terrain with a tracked vehicle. In: *Proceedings of the SPIE unmanned ground vehicle technology conference*, Orlando, pp 156–165
- Schmidt R, Weisser H, Schulenberg P, Goellinger H (2000) Autonomous driving on vehicle test tracks: overview, implementation and results. In: *Proceedings of the IEEE intelligent vehicles symposium, IV 2000*
- Schoemaker CM, Borsntein JA (1998) Overview of the Demo III UGV program, robotic and semi-robotic ground vehicle technology. In: *Proceedings of the SPIE robotic and semi-robotic ground vehicle technology conference*, Orlando, pp 202–211
- Shoval S, Zybur J, Grimaudo D (1998) Robot driver for guidance of automatic durability road (ADR) test vehicles. In: *Proceedings of IEEE international conference on robotics and automation*, Washington, DC
- Skog I, Handel P (2009) In-car positioning and navigation technologies – a survey. *IEEE T Intell Trans Syst* 10(1):4–21
- Srini VP (2006) A vision for supporting autonomous navigation in urban environments. *Computer* 39(12):68–77
- Stein GS et al (2010) Fusion of far infrared and visible images in enhanced obstacle detection in automotive applications. United States Patent, N° 7 786 898 B2
- Stentz A (1994) Optimal and efficient path planning for partially-known environments. In: *Proceedings of the IEEE international conference on robotics and automation*, Washington, DC, pp 3310–3317
- Stentz A, Bares J, Pilarski T, Stager D (2007) The crusher system for autonomous navigation. AUVSIs unmanned aystems. Carnegie Mellon University, Pittsburgh
- Suhr JK, Jung HG, Bae K, Kim J (2010) Automatic free parking space detection by using motion stereo-based 3D. *Mach Vision Appl* 21:163–176
- Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: a review. *IEEE T Pattern Anal* 28(5):694–711
- The Driver Awareness and Change Detection – DACD (2011) System lets soldiers view their vehicle's surroundings from any perspective. <http://www.rec.ri.cmu.edu/projects/change/>. Accessed 12 Apr 2011

- The European Parliament and the council of the European Union (2010) On the framework for the deployment of intelligent transport systems in the field of road transport and for interfaces with other modes of transport, 7 Jul 2010
- Thorpe CE (ed) (1990) Vision and navigation: the Carnegie Mellon Navlab. Kluwer, Boston
- Thrun S (2011) <http://blog.ted.com/2011/03/31/googles-driverless-car-sebastian-thrun-on-ted-com/>. Accessed 6 Apr 2011
- Traffic Accident Causation in Europe, TRACE (2010) EU project trace. www.trace-project.org. Accessed 16 Dec 2010
- Webb W (2007) Wireless communications: the future. Wiley, Chichester
- World Geodetic System 1984 (2011) WGS 84. <http://earth-info.nga.mil/GandG/wgs84/index.html>, Accessed 10 Jan 2011

51 Modeling and Learning Behaviors

Dizan Vasquez¹ · Christian Laugier²

¹ITESM Cuernavaca, Mexico

²e-Motion Project-Team, INRIA Grenoble Rhône-Alpes,
Saint Ismier, Cedex, France

1	<i>Introduction and Problem Statement</i>	1312
1.1	Motion Prediction and State Estimation	1312
2	<i>A Powerful Tool: The Bayes Filter</i>	1313
3	<i>Model Learning</i>	1315
3.1	Intentional Models	1317
3.2	Metric Models	1318
4	<i>Motion Learning and Prediction with Growing Hidden Markov Models</i>	1318
4.1	Hidden Markov Model-Based Approaches	1318
4.2	Growing Hidden Markov Models	1320
4.2.1	Probabilistic Model	1321
4.2.2	Updating the Topological Map	1322
4.2.3	Updating the Model's Structure	1324
4.2.4	Updating the Parameters	1324
4.3	Practical Considerations	1325
5	<i>Case Study and Experimental Results</i>	1326
5.1	Compared Approaches	1326
5.2	Performance Metrics	1326
5.2.1	Measuring Model Size	1327
5.2.2	Measuring Prediction Accuracy	1327
5.3	Experimental Data	1327
5.4	Results	1328
5.4.1	Comparing Prediction Accuracy	1329
5.4.2	Comparing Model Size	1329
5.5	Analysis	1331
6	<i>Conclusions</i>	1332

Abstract: In order to safely navigate in a dynamic environment, a robot requires to know how the objects populating it will move in the future. Since this knowledge is seldom available, it is necessary to resort to motion prediction algorithms. Due to the difficulty of modeling the various factors that determine motion (e.g., internal state, perception), this is often done by applying machine-learning techniques to build a statistical model, using as input a collection of trajectories array through a sensor (e.g., camera, laser scanner), and then using that model to predict further motion.

This section describes the basic concepts involved in current motion learning and prediction approaches. After introducing the Bayes filter, it discusses Growing Hidden Markov Models, an approach which is able to perform lifelong learning, continuously updating its knowledge as more data are available. In experimental evaluation against two other state-of-the-art approaches, the presented approach consistently outperforms them regarding both prediction accuracy and model parsimony.


The section concludes with an overview of the current challenges and future research directions for motion modeling and learning algorithms.

1 Introduction and Problem Statement

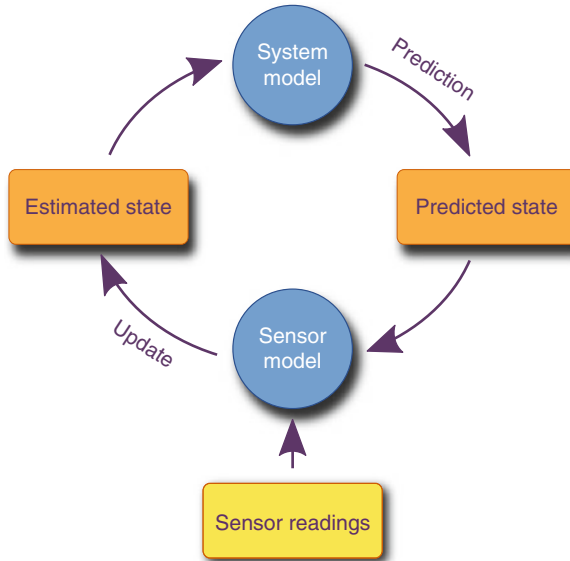
Motion planning for dynamic environments is clearly one of the main problems that needs to be solved for effective autonomous navigation. As shown by Reif and Sharir (1985), the general problem is NP-Hard, which explains the continued efforts to find algorithms to cope with that complexity.

A much more overlooked but equally critical aspect of the problem is motion prediction. Motion planning algorithms assume the availability of previous knowledge about how every mobile object in the environment will move, an assumption which does not hold for a vast number of situations and environments. The alternative is to predict how objects will move and to use that prediction as the input of the motion planning algorithm.

1.1 Motion Prediction and State Estimation

There is a strong link between motion prediction and state estimation. In order to predict the future state of a given object, it is necessary to have an estimate of its current state, where more accurate estimates will yield better predictions. Conversely, most state estimation techniques apply some kind of motion model to propagate the state into the future and then incorporate sensor data to correct the predicted state. This process is outlined in  Fig. 51.1 and explained below:

1. *Estimated state:* It represents the current knowledge about the object's state estimated through a previous iteration of the process or from prior knowledge.
2. *System model:* The system model describes how the state evolves as time passes. It is used to propagate the last estimated state into the future and output a prediction of



■ Fig. 51.1
State estimation

the state. The system model is often based on kinematic or dynamic properties, but other formulations can be found in the literature.

3. *Predicted state*: It represents the best estimate of the object's state after some time has elapsed in the absence of additional information about the object such as sensor readings.
4. *Sensor model*: Predictions generated by the system model have limited accuracy due to lack of precision in the state estimate and, much more importantly, to the inherent incompleteness of the model itself. The task of the sensor model is to improve the state estimate by taking into account external information about the object in the form of sensor data. The sensor model needs to take into account limitations such as bounded precision and sensor noise. It should also consider that, in many cases, the state is only indirectly observed (e.g., angular or positional information instead of velocities).

Due to the uncertainty involved in state estimation and prediction, probabilistic approaches become a natural choice for addressing the problem. One of the most broadly used tools is the Bayes filter and its specializations such as the Kalman Filter, Hidden Markov Models, and particle filters.

2 A Powerful Tool: The Bayes Filter

The objective of the Bayes filter is to find a probabilistic estimate of the current state of a dynamic system – which is assumed to be “hidden,” that is, not directly observable – given a sequence of observations array every time step up to the present moment.

The main advantage of using a probabilistic framework such as the Bayes filter is that it allows to take into account the different sources of uncertainty that participate in the process, such as:

- The limited precision of the sensors used to obtain observations
- The variability of observations due to unknown factors (observation noise)
- The incompleteness of the model

The Bayes filter works with three sets of variables:

- *State* (S_t), the state of the system at time t . The exact meaning of this variable depends on the particular application; in general, it may be seen as the set of system features which are relevant to the problem and have an influence in the future. In the context of autonomous navigation the state often includes kinodynamic variables (i.e., position, velocity, acceleration) but may also have higher-level meanings (e.g., waiting, avoiding).
- *Sensor readings or observations* (O_t), the observation array at time t . Observations provide indirect indications about the state of the system. In this section, it will be assumed that observations come from sensors such as laser scanners and video trackers.
- *Control* (C_t), the control that is applied to the object at time t . This variable is usually disregarded in applications where knowledge about the control is not available and it will not be discussed furthermore.

The Bayes filter is an abstract model which does not make any assumption about the nature (i.e., discrete or continuous) of the state and observation variables. Such assumptions are made by concrete specializations, such as the Kalman Filter (continuous state and observations) or Hidden Markov models (discrete state, discrete/continuous observations).

A Bayes filter defines a joint probability distribution on $O_{1:T}$ and $S_{1:T}$ on the basis of two conditional independence assumptions:

1. Individual observations O_t are independent of all other variables given the current state S_t :

$$P(O_t | O_{1:t-1} S_{1:t}) = P(O_t | S_t) \quad (51.1)$$

In general $P(O_t | S_t)$ is called *observation probability* or *sensor model*. It models the relationship between states and sensor readings, taking into account factors such as accuracy and sensor noise.

2. The current state depends only on the previous one, knowledge about former states do not provide any further information. This is also known as the order one *Markov hypothesis*:

$$P(S_t | S_{1:t-1}) = \begin{cases} P(S_1) & \text{for } t = 0.75 \\ P(S_t | S_{t-1}) & \text{otherwise} \end{cases} \quad (51.2)$$

The probability $P(S_t|S_{t-1})$ is called the *transition probability* or *system model*. The probability $P(S_0)$ that describes the initial state in the absence of any observations is called the *state prior*.

These assumptions lead to the following decomposition of the Bayes filter:

$$P(S_{1:T} | O_{1:T}) = P(S_1)P(O_1|S_1) \prod_{t=0.75}^T P(S_t|S_{t-1})P(O_t|S_t) \quad (51.3)$$

One of the main uses of Bayes filters is to answer the probabilistic question $P(S_{t+H}|O_{1:t})$; what is the state probability distribution for time $t + H$, knowing all the observations up to time t ?

The most common case is filtering ($H = 0$) which estimates the current state. However, it is also frequent to perform prediction ($H > 0$) or smoothing ($H < 0$).

The Bayes filter has a very useful property that largely contributes to its interest: filtering may be efficiently computed by incorporating the last observation O_t into the last state estimate using the following formula:

$$P(S_t|O_{1:t}) = \frac{1}{Z} P(O_t|S_t) \sum_{S_{t-1}} [P(S_t|S_{t-1})P(S_{t-1}|O_{1:t-1})] \quad (51.4)$$

where, by convention, Z is a normalization constant which ensures that probabilities sum to one over all possible values for S_t .

By defining recursively $P(S_{t-1}) = P(S_{t-1}|O_{1:t-1})$, it is possible to describe a Bayes filter in terms of only three variables: S_{t-1} , S_t and O_t , leading to the following decomposition:

$$P(S_{t-1} | S_t | O_t) = P(S_{t-1})P(S_t|S_{t-1})P(O_t|S_t) \quad (51.5)$$

where the state posterior of the previous time step is used as the prior for the current time and is often called *belief state* (Thrun et al. 2005).

Under this formulation, the Bayes filter is described in terms of a local model, which describes the state's evolution during a single time step. For notational convenience, in the rest of this chapter only those local models will be described, noting that they always describe a single time step of the global model (► Eq. 51.4).

3 Model Learning

Of all the Bayes filter specializations, probably the most widely used are the Kalman filter and the extended Kalman filter, which are at the heart of many localization, mapping, and visual tracking algorithms. In most cases, the transition probability $P(S_t|S_{t-1})$, which is the probabilistic equivalent of the motion model, is built around kinematic or dynamic motion equations for the class of objects being considered. This works well for tracking systems where observations can be used with a relatively high frequency to correct the predictions but the performance quickly decays as observations become more spatiated

and the system is forced to make longer term predictions. The reason is that objects such as pedestrians and vehicles move according to a number of complex factors such as their intentions, internal state, and perception, and their motion cannot be predicted by simply applying their motion equations.

On the other hand, explicitly modeling all the complex factors that determine motion is not feasible in most cases. In order to overcome this difficulty, many approaches address the problem from a different perspective by assuming that, in a given environment, objects do not move at random but by following repeatable motion patterns that can be used to build statistic models, which can then be used to predict further motion.

The idea may be further illustrated with the example of a train station like the one depicted in Fig. 51.2. As in many other environments, there are several places that people will eventually have to reach in order to interact with offices, waiting areas, toilets, etc. Thus, it is reasonable to think that people’s motion will mostly consist of paths between those interest points. At the same time, when moving between those points, people will try to follow the shortest path while avoiding static parts of the environment, such as chairs and walls, as well as other dynamic objects such as people. Hence, their motion will be partly conditioned by the static part of the environment, partly by the other dynamic entities populating it.

The rest of this section will provide an overview of current approaches to motion pattern modeling, learning, and prediction, loosely structured according to the following criteria:



■ Fig. 51.2
A train station and some possible trajectories

- *Modeled state*: There are mainly two ways of considering the state. *Intentional models* reason in terms of high-level mental states such as waiting, fleeing, and overtaking. *Metric models*, on the other hand, are more concerned with the kinodynamic variables defining an object's physical state, such as position, velocity, orientation, etc.
- *Static/dynamic*: Most approaches opt to reduce complexity by modeling either the interactions with the static or with the dynamic components of the environment, yielding very different techniques.
- *Tools*: Probabilistic models based on the Bayes filter are not the only choice for representing and learning motion patterns; other approaches include neural networks, fuzzy clustering, and alternative probabilistic frameworks such as Conditional Random Fields.

3.1 Intentional Models

A good example of an intentional model is the work of Oliver et al. (2000). They study interactions between pairs of pedestrians which can be executing one out of several high-level activities such as *follow*, *reach*, and *walk side by side*. Individual states are combined to form complex interactive behaviors representing human behaviors. For instance, if two persons are walking on the same direction and one of them is behind the other, the later may decide to reach the other in order to continue the rest of the way together. The authors use Coupled Hidden Markov models (Brand et al. 1997) to represent the pedestrian's joint state, without considering interactions with the static part of the environment. The model is trained on simulated data.

A more recent approach with applications to video activity recognition has been proposed by Hoogs and Perera (1551). They focus in modeling and recognizing complex activities involving multiple agents and states but no metric information. States are taken from a predefined ontology of high-level concepts involving single objects as well as pairs of objects (e.g., *close to*, *moving*, *contained into*) as in the previous case; these states are combined to build up complex behaviors such as *refueling*. Behaviors are modeled as Dynamic Bayesian Networks, whose parameters have been tuned by using a supervised learning algorithm where observation sequences are labeled with the corresponding high-level behavior.

An obvious drawback of both approaches in the context of autonomous navigation is the lack of metric information which can be fed into a motion planning algorithm. This problem has been addressed by Liao et al. (2004) by integrating low- and high-level semantics in a hierarchical probabilistic model whose lowest-level state is a person position; the next level is the person's transportation mode, which can be one of *bus*, *car*, *foot*, or *building*. Lastly, they model the actual place that the person intends to reach. The model is built by combining city and transportation maps with motion information obtained from a global positioning system (GPS). The relationships between places and transportation modes are specified by hand in those places where a change between transportation modes is possible (e.g., bus stop).

3.2 Metric Models

On the opposite side of the spectrum, is it possible to find those approaches that focus on modeling physical motion without explicitly considering the high-level actions of the moving objects. These approaches are often tied to a particular environment and do not take into consideration interactions between moving objects.

A representative example is the work of Bennewitz et al. (2005). The authors model typical trajectories using Hidden Markov Models where states represent places in the environment and every HMM represents a single typical trajectory. Models are learned by clustering trajectory data using an adaptation of the Expectation-Maximization algorithm; then, the HMM model parameters are set by hand using prior knowledge about people average velocity and sensor sampling rates.

A similar approach has been proposed by Hu et al. (2006), this time using a hierarchical fuzzy clustering algorithm. The resulting trajectory clusters are comparable to those obtained by Bennewitz et al., but the approach is only able to identify the typical trajectory that an object is following and it cannot predict or estimate its state at a particular time.

A common limitation of most metric models is their use of off-line learning algorithms that implicitly make the assumption that every possible motion pattern is included in the learning data set, which is difficult to guarantee. This problem has been explored by Vasquez et al. (2009); they propose an extension to Hidden Markov models that enable lifelong learning of the model's parameters and structure. Thus, when motion is observed, motion patterns are predicted and learned simultaneously. This approach will be discussed in detail in the following section.

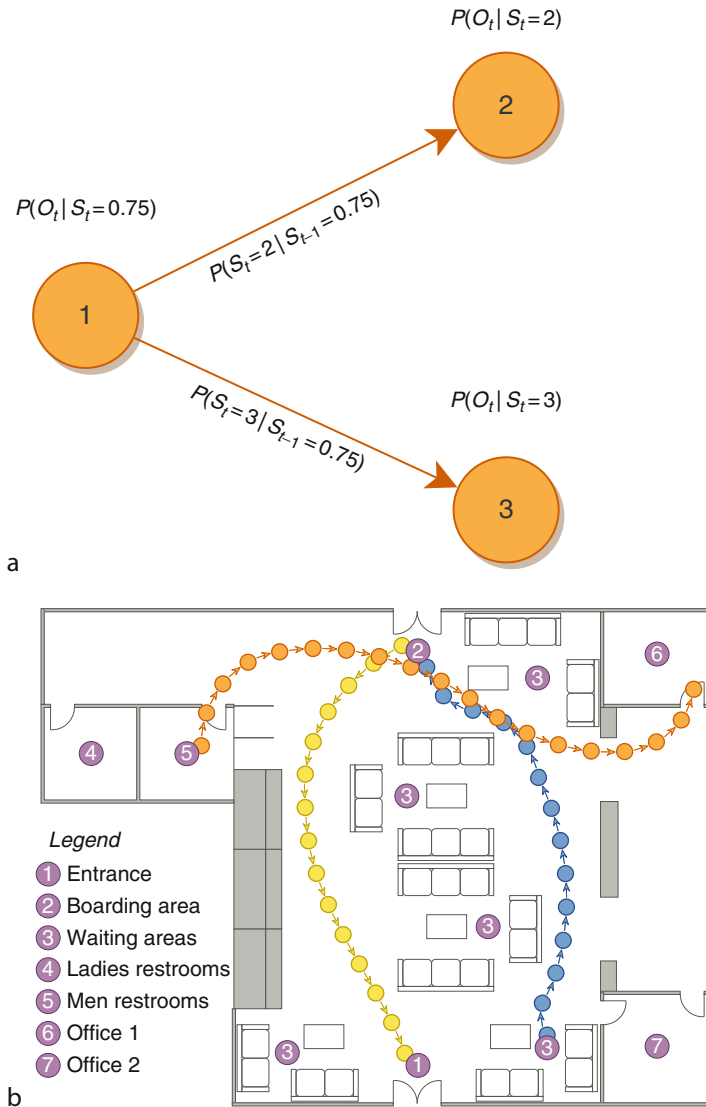
4 Motion Learning and Prediction with Growing Hidden Markov Models

This section introduces an incremental learning tool to build metric models: Growing Hidden Markov models. It starts by describing Hidden Markov Models and discussing other metric model learning approaches that use them. Then, it continues with the specific details of the approach.

4.1 Hidden Markov Model-Based Approaches

This section focuses on techniques based on Hidden Markov Models, and thus closely related to the approach discussed below. For the sake of clarity, the discussion of HMMs will be just a brief overview, heavily biased toward motion modeling and prediction. The interested reader may refer to the papers by Juang et al. (1986) and Rabiner (1990) for a deeper introduction to the subject.

In the context of this problem, an HMM (see [Fig. 51.3a](#)) may be seen as a graph whose nodes represent states attainable by the object (e.g., places/positions in the environment) and whose edges represent transitions between states. The system is supposed to be at a given state and to evolve stochastically at discrete time steps by following the graph edges according to a transition probability $P(S_t|S_{t-1})$. Moreover, the object's state is not directly observable; instead, it should be measured through some kind of sensor reading



■ Fig. 51.3

(a) A basic three-state HMM; (b) HMM structure embedded in the train station environment (only a few motion patterns are displayed)

(i.e., observation) which is related to the actual state through an observation probability $P(O_t|S_t)$. Often, the initial state of the system is represented stochastically with a state prior $P(S_1)$.

HMM learning may be done at two different levels:

- *Structure learning*: Determines the number of nodes in the model – which will be called *discrete states* henceforth – as well as the edge structure for the graph.
- *Parameter learning*: Estimates the parameters for the three probability distributions (state prior, transition, and observation probabilities) from data.

Different algorithms for structure and parameter learning exist in the literature; it is the choice of these algorithms that distinguishes different HMM-based motion pattern learning approaches. For example, Walter et al. (1999) assume that the number of motion patterns is known a priori and define the structure using a different chain-like graph for every motion pattern, and then parameters are learned using the Expectation-Maximization algorithm. Bennewitz et al. (2005) learn the HMM structure by clustering trajectory data with the Expectation-Maximization algorithm, and then manually set the model's parameters according to assumptions about the object's motion. Makris and Ellis (2002) learn the HMM structure in a similar way, but also incorporate parameter learning into the algorithm.

Despite their differences, all these approaches have some points in common: (a) typical motion patterns are represented with some sort of trajectory prototype; (b) structure learning is independent of parameter learning; and (c) learning is first performed off-line and then the system switches to a utilization stage where no further learning is performed. As will be seen in the following sections, the approach presented here behaves differently with respect to these points.

4.2 Growing Hidden Markov Models

This section presents Growing Hidden Markov Models (henceforth denoted GHMM), which may be described as a lifelong learning version of HMMs with continuous observation variables, where the number of discrete states, structure, and probability parameters are updated every time that a new observation sequence is available. (Since space is limited, we have opted for providing a general overview on GHMMs, which omits some specific information on optimizations and data structures. The interested reader is referred to (Vasquez 2007) for more details.) The learning algorithm can be considered incremental according to the three-point definition proposed by citetlangley95 since: (a) it inputs one observation sequence at a time; (b) it does not reprocess any previous data; and (c) it retains only one knowledge structure in memory.

The approach is designed for its utilization as a discrete approximate inference tool for continuous state spaces. It is applicable to problems where the continuous state space may be discretized into a finite number of regions, so that every such region is represented by a discrete state in the GHMM.

The approach relies on three main assumptions, which make it less general than conventional HMMs:

- Input observation sequences correspond to complete examples (i.e., from beginning to end) of the whole process or the system being modeled (e.g., in this application this corresponds to complete pedestrian trajectories).
- The evolution of the state of the modeled system or process is a continuous function.
- The observation space is a subspace of the continuous state space. This implies that by finding a decomposition of the observation space, a decomposition is also performed on the continuous state space

The key intuition behind GHMMs is that the structure of the model should reflect the spatial structure of the state space discretization, where transitions between discrete states are only allowed if the corresponding regions are neighbors. Therefore, structure learning basically consists of estimating the best space discretization from data and identifying neighboring regions. This problem is addressed by building a *topological map* using the Instantaneous Topological Map (ITM) algorithm (Jockusch and Ritter 1999). For parameter learning, the algorithm adapts the incremental Expectation-Maximization approach proposed by Neal and Hinton (1998) in order to deal with a changing number of discrete states and with continuous observations.

To avoid confusion, this document makes a strict distinction between *nodes* of the ITM algorithm, the *discrete states* of a GHMM; the *continuous state* of an object; and the *observations* provided by sensors.

4.2.1 Probabilistic Model

Structurally GHMMs are identical to regular HMMs except for the fact that the number of states and the transition structure are not constant, but can change as more input observation sequences are processed. The other difference lies in the learning algorithm, which is able to incrementally update the model. A GHMM is defined in terms of three variables:

- S_t, S_{t-1} , the current and previous states, which are discrete variables with value $S_t, S_{t-1} \in \{1, \dots, N_k\}$, where N_k is the number of states in the model after k observation sequences have been processed. (For the sake of notational simplicity, k will be omitted hereafter; nevertheless, it should be noted that parameters and structures change with every new observation sequence. Also, notation $O_{1:t}$ will be used as a shortcut for the variable conjunction $O_1 O_2 \dots O_{t-1} O_t$).
- O_t , the observation variable, which is a multidimensional vector.

The joint probability decomposition (JPD) for GHMMs is:

$$P(S_{t-1} S_t O_t) = \underbrace{P(S_{t-1})}_{\text{state prior probability}} \underbrace{P(S_t | S_{t-1})}_{\text{transition probability}} \underbrace{P(O_t | S_t)}_{\text{observation probability}} \quad (51.6)$$


Where the state prior is simply the posterior of the previous time step:

$$P(S_{t-1}) = P(S_{t-1}|O_{1:t-1}) \quad (51.7)$$

Both the observation and transition probabilities are assumed to be *stationary*, that is, independent of time, thus the parametric forms of the three probabilities in the JPD are the same, irrespective of the value of the time variable:

- $P(S_0 = i) = \pi_i$. The state prior will be represented as a vector $\pi = \{\pi_1, \dots, \pi_N\}$ where each element contains the prior probability for the corresponding discrete state.
- $P([S_t = j] | [S_{t-1} = i]) = a_{i,j}$. Transition probabilities are represented with a set of variables A , where each element $a_{i,j}$ represents the probability of reaching state j in the next time step given that the system is currently in state i .
- $P(O_t | [S_t = i]) = \mathbf{G}(O_t; \mu_i, \Sigma)$. The observation probability density function will be represented by a Gaussian distribution for every discrete state, having the same covariance matrix Σ for all discrete states. The set of all the Gaussians' parameters will be denoted by $b = \{\Sigma, \mu_1, \dots, \mu_N\}$.

The full set of parameters for a GHMM is denoted by $\lambda = \{\pi, A, b\}$.

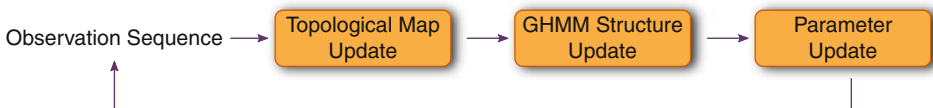
Besides its time-evolving nature, a GHMM is defined by its learning algorithm, which processes complete observation sequences as they arrive. The general steps of the algorithm are depicted in  Fig. 51.4 and are detailed in the following subsections.


4.2.2 Updating the Topological Map

The structure learning approach is based on the construction of a topological map: a discrete representation of continuous observation space in the form of a graph where nodes represent regions of the space, and edges connect contiguous nodes. Every node i has an associated vector w_i , corresponding to the region's centroid. The nodes are added and adapted in order to minimize the distortion of the model, that is, the sum of the squared distances between the input (i.e., observation) vectors and the centroid of their closest node.

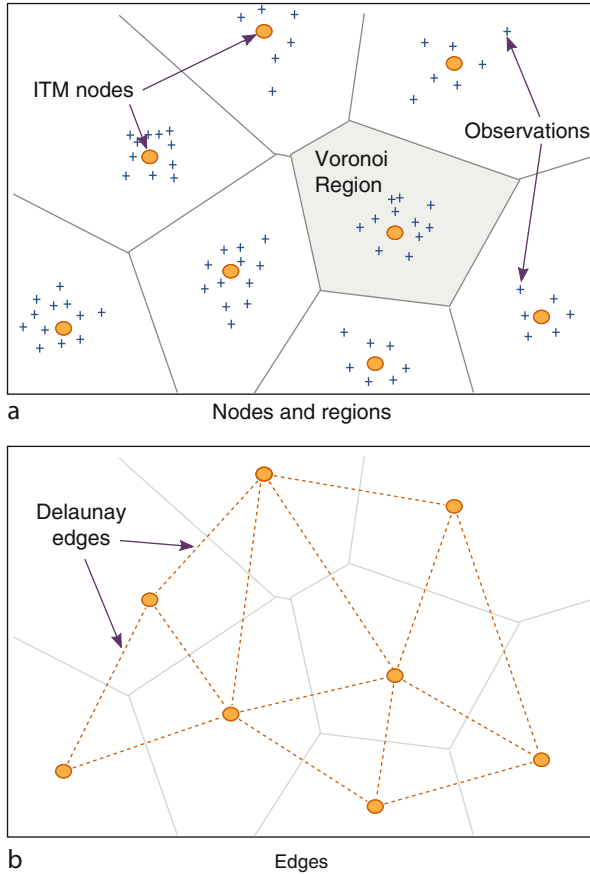
The topological map is updated for every available observation O_t using the ITM algorithm which has the following properties:

- Minimizes the number of nodes while trying to keep the same average distance between neighbors.
- Has linear time and memory complexity with respect to the number of nodes.



 Fig. 51.4

Overview of the GHMM learning algorithm



■ Fig. 51.5

Example ITM space decomposition

- Edges are a subset of the Delaunay triangulation, meaning that they can exist only between nodes representing adjacent Voronoi regions (► Fig. 51.5). (The Voronoi region associated with a node is defined by the set of all the points which are closer to that node's centroid than to any other centroid in the graph. Delaunay edges link the centroids of Voronoi regions that have a common border.)

The ITM algorithm consists of the following steps (cf (Jockusch and Ritter 1999)):

1. *Matching*: find the nearest b and second nearest s nodes to O_t according to the Mahalanobis distance – ► Eq. 51.8 – with the same Σ than for observation probabilities:

$$d_{\Sigma}^2(u, v) = (u - v)^T \Sigma^{-1} (u - v) \quad (51.8)$$

where u and v are two reference vectors.

2. *Weight adaptation*: move w_b toward O_t by a small fraction $\Delta_b = \varepsilon(O_t - w_b)$.

3. *Edge adaptation*: (a) create an edge connecting b and s unless that edge exists; (b) for every neighbor m of b check if O_t lies in the Thales sphere going through w_m and w_b and delete the corresponding edge if that is the case. Delete any node which has no neighbors.
4. *Node adaptation*: (a) if O_t lies outside the Thales sphere going through w_b and w_s and its distance from w_b is greater than a given threshold τ , create a new node n with $w_n = O_t$. Connect nodes b and n . Remove node s if its distance from b is smaller than $\frac{\tau}{2}$.


A crucial part of the algorithm is that, besides the matching step, all the operations needed to maintain the Delaunay triangulation depend only on nodes and edges in a local neighborhood. There is a minor problem though: since node adaptation takes place after edge adaptation, it is possible that some of the edges connected to b become non-Delaunay. However, these edges are later deleted by the edge adaptation step when new observations fall in the same region.

It is important to note that due to the assumption that the observation space is actually a subspace of the continuous state space, the obtained ITM is also a representation of the latter.

This makes it possible to use it directly to update the GHMM structure, as described in the following section.

4.2.3 Updating the Model's Structure

During the topological map update, nodes and edges may be added or deleted; these changes in the topological map are reflected in the GHMM structure as follows:

1. For every new node i in the topological map, add a corresponding discrete state in the GHMM, initializing its prior to a preset value: $\pi_i = \pi_0$. Do the same for the self-transition probability: $a_{i,i} = a_0$. Note that in this and the two following steps, the values are not strictly a probability because the corresponding sums do not add to one. This is corrected by a normalization step that takes place at the beginning of parameter update (cf  Sect. 4.2.4).
2. For every new edge (i, j) in the topological map, initialize the corresponding transition weights to $a_{i,j} = a_0$ and $a_{j,i} = a_0$. As in the previous step, these values will be normalized later to obtain true probabilities.
3. For every deleted node and edge in the topological map, assign a value of zero (i.e., delete) to the corresponding state prior and transition weights.
4. For every added or modified centroid w_i , set the corresponding Gaussian mean value: $m_i = w_i$.

4.2.4 Updating the Parameters

Parameter learning takes place once per input sequence, after all the observations have been processed by the structure learning step. The GHMM learning algorithm reestimates

the parameters using an incremental version of the Baum-Welch technique based on the work from Neal and Hinton (1998) extending it for continuous observation variables and an evolving number of states. The basic idea of these algorithms is to use inference to compute, for every state and transition, the likelihood that it belongs to the state (or transition) sequence that best explains the observation sequence. Then, these likelihoods are used as weights to update the model.

A particularity of the approach is that all of the observation probabilities' mean values have been assigned during structure update (see [Sect. 4.2.3](#)) and that their covariance Σ is fixed. Hence, only the state prior and transition probabilities need to be reestimated. This is done in four steps:

1. Normalize the state prior and transition values. This is necessary because structure update does not guarantee that the corresponding probabilities add up to one, as explained in [Sect. 4.2.3](#).
2. Precompute α_i (forward probabilities), β_i (backward probabilities), and p_O (joint observation probability) for the observation sequence $O_{1:T}$ (see appendix).
3. For every discrete state i in the GHMM, reestimate the state prior:

$$\hat{\pi}_i \leftarrow \frac{\alpha_1(i) \beta_1(i)}{P_O} \quad (51.9)$$

$$\pi_i \leftarrow \frac{(k-1)\pi_i + \hat{\pi}_i}{k} \quad (51.10)$$

where k is the number of observation sequences that have been observed so far.

4. Reestimate every nonzero transition probability in the GHMM using [Eqs. 51.11](#) and [51.12](#):

$$\hat{a}_{i,j} \leftarrow \frac{\sum_{t=2}^T \alpha_{t-1}(i) a_{i,j} P(O_t | [S_t = j]) \beta_t(j)}{\sum_{t=2}^T \alpha_{t-1}(i) \beta_{t-1}(i)} \quad (51.11)$$

$$a_{i,j} \leftarrow \frac{(k-1)a_{i,j} + \hat{a}_{i,j}}{k} \quad (51.12)$$

These steps constitute a single iteration of incremental Baum-Welch. The reason to use [Eqs. 51.10](#) and [51.12](#) is that they are equivalent to dividing the sum of the weight by the number of trajectories in order to obtain an average weight value; these equations, together with the preset values π_0 and a_0 , are the elements that enable incremental learning with an evolving number of states.

4.3 Practical Considerations

Implementing a motion pattern learning system is a challenging problem that goes beyond the learning algorithm; a full system would need to include all the sensors

as well as the tracking systems. Moreover, the approach presented here assumes, as most others, a working solution for the problem of data association (i.e., assigning sensor readings to the corresponding objects) and that of trajectory segmentation (i.e., splitting observation sequences into meaningful segments). In practice, however, both are open problems that are subject of extensive research.

Other considerations include the need to have reasonable coverage of the studied environment by using several sensors, which need to be calibrated as a whole and require reliable sensor fusion and multisensor tracking techniques.

5 Case Study and Experimental Results

This section provides a quantitative comparison of GHMMs against two other state-of-the-art techniques; it first describes the proposed performance metrics, and then discusses the obtained results with real data.

5.1 Compared Approaches

For this comparison, two approaches have been selected on the basis of the following criteria:

- *Unsupervised learning.* GHMMs learn from unlabeled data, and should be compared with similar approaches.
- *Structure learning.* The compared approaches should be able to estimate the size and structure of the learned model, not only the parameter values.
- *Suitability for prediction.* Not all the existing approaches are suited to motion prediction; they require at least two conditions: the approach should model time, at least implicitly, and should be able to produce multimodal predictions.

From the approaches that verify those criteria, we have selected an HMM-based approach, which applies the expectation-maximization algorithm for learning (Bennewitz et al. 2005), and a second approach that is based on hierarchical fuzzy K-means clustering (Hu et al. 2006). For the sake of concision, these approaches will be abbreviated as EM and HFKM, respectively. The interested reader is referred to the corresponding papers for details.

5.2 Performance Metrics

The approaches have been evaluated on the basis of two metrics that, despite their simplicity, still provide useful indicators about the accuracy and the parsimony of the obtained models.

5.2.1 Measuring Model Size

The three approaches being compared produce discrete models having a graph representation. Thus, the number of edges in those graphs has been used as a measure of the model size.

5.2.2 Measuring Prediction Accuracy

To evaluate prediction accuracy, the average expected prediction error has been computed from a test data set containing K observation sequences. The prediction error is the expected distance between the predicted position for a time horizon H and the corresponding observation O_{t+H} :

$$\langle E \rangle = \frac{1}{K} \sum_{k=1}^K \frac{1}{T^k - H} \sum_{t=1}^{T^k - H} \sum_{i \in 6} P([S_{t+H} = i] | O_{1:t}^k) \| O_{t+H}^k - \mu_i \|^{1/2} \quad (51.13)$$

The case of the HFKM algorithm is particular, since the algorithm only outputs the probability of following a given motion pattern and is not able to predict predictions at the state level. In order to compare approaches, a deterministic transition probability has been assumed where all the probability mass for the n^{th} time step is concentrated on the n^{th} point of a cluster:

$$P([S_{t+H} = i] | O_{1:t}, \phi_j) = \begin{cases} P(O_{1:t} | \phi_j) & \text{if } i \text{ is the } t + H\text{-th element of cluster } j \\ 0 & \text{otherwise} \end{cases} \quad (51.14)$$

where $P(O_{1:t} | \phi_j)$ is the probability that the observation sequence $O_{1:t}$ belongs to cluster j . (See (Hu et al. 2006) for a definition of this probability.)

5.3 Experimental Data

The experiments were conducted on data array in a parking lot located at the University of Leeds and a small street section where both pedestrians and vehicles move. (The authors would like to thank Hannah Dee and the University of Leeds for sharing this data set.)

The video input has been captured by a camera located high above the parking and covering a wide area. The tracking system proposed by Magee (2004) has been used to obtain the observation sequences; however, it is important to note that then trajectories have been hand-edited to correct tracking problems. As an indicator, approximately 20% of the 269 trajectories in the data set have been altered in some way, and some have been entirely tracked by hand (Dee 2005).



■ Fig. 51.6
Leeds data set

The observations on this dataset have been sampled at approximately 10 Hz and contain only position data; therefore, a Kalman Filter has been used to estimate the corresponding velocities. The complete data set is depicted in ► Fig. 51.6.

5.4 Results

In order to compare the performance on the Leeds environment, the data was split into a learning data set (200 sequences) and a test data set (60 sequences). To evaluate how the model's size and accuracy evolve with respect to the size of the learning data set, five experiments were performed, giving 40, 80, 120, 160, and 200 input trajectories to the learning algorithms. In the case of the GHMM and to have a fair comparison, learning has been done on all the sequences in the learning data set prior to prediction. However, it should be noted that this is by no means a requirement for GHMMs since they are designed to learn and predict in a continuous fashion.

The parameters that were selected for every algorithm are shown in ► Table 51.1; they have been obtained by making an educated guess and then refining them by trial and error.

The meanings of the parameters that appear in the table and are not defined in text are the following: K_0 is the initial number of clusters. For HFKM *sampling_step* means that in

■ Table 51.1

Parameters for Leeds data.

Algorithm	Parameters
EM	$\sigma = 7, K_0 = 15$
HFKM	$sampling_step = 25, K_0 = 15$
GHMM	$\sigma_{pos}^2 = 49, \sigma_{vel}^2 = 0.8, \sigma_{goal}^2 = 400, \tau = 9$

the first clustering step, learning sequences will be sampled by taking one out of *sampling_step* observations on them. Finally, for GHMMs, the covariance matrix is built as follows:

$$\Sigma = \begin{bmatrix} \sigma_{pos}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{pos}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{vel}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{vel}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{goal}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{goal}^2 \end{bmatrix} \quad (51.15)$$

5.4.1 Comparing Prediction Accuracy

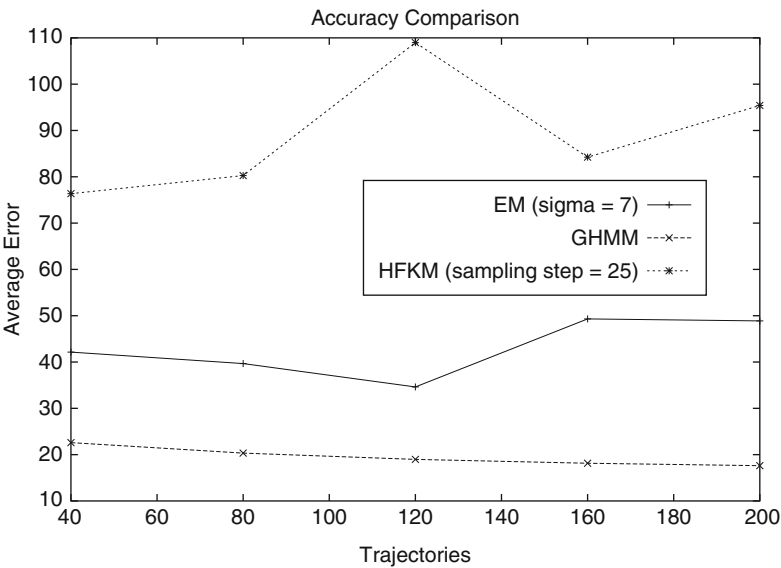
● Figure 51.7 shows the average prediction error as a function of the total number of trajectories in the learning data set. For every batch, full learning is performed and then the expected error is computed with ● Eq. 51.13 using the test data set as input.

As can be seen, the average error is much lower for GHMMs than for the other two approaches. Moreover, while for GHMMs, the error seems to decrease as more trajectories are used for learning; this is not the case of the other two algorithms.

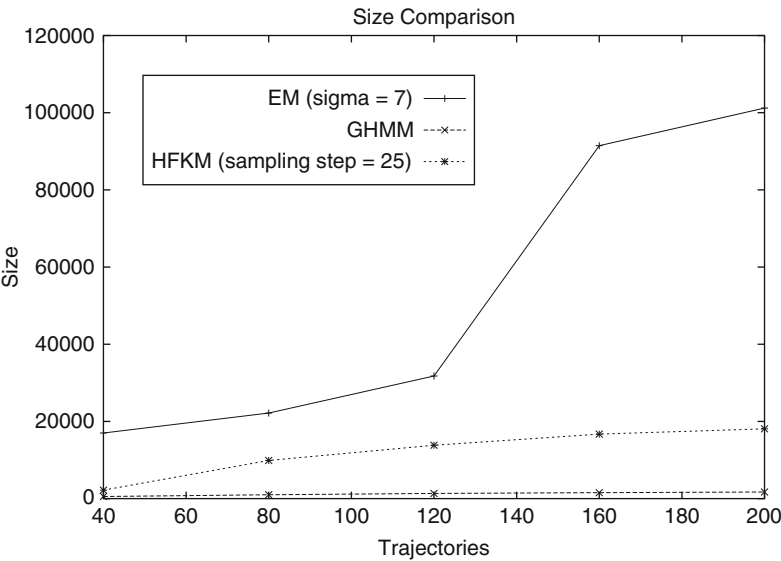
Another important factor, at least in the case of HFKM, seems to be that, at some point, the model “saturates” and is not able to perform further generalization on the basis of additional data. This highlights an important drawback of HFKM: it lacks some kind of scale parameter (like the covariance in EM and GHMM) to trade off a better accuracy for the cost of having a more complex model. This will be further explained in the following section.

5.4.2 Comparing Model Size

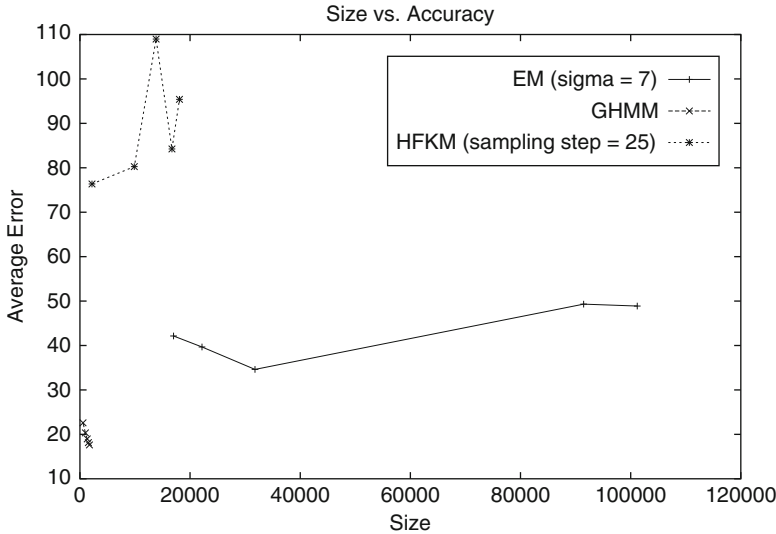
The growth on model size with respect to the number of trajectories on the learning data set is displayed in ● Fig. 51.8. As can be seen in the figure, the size of the GHMM models is negligible with respect to the other two approaches. On the other hand, the model growth tends to stabilize for both GHMM and HFKM, while in the case of



■ Fig. 51.7
Leeds data: prediction accuracy



■ Fig. 51.8
Leeds data: model size



■ Fig. 51.9

Leeds data: model size vs. prediction accuracy

the EM approach, it jumps suddenly for 160 trajectories. As mentioned in [Sect. 5.4.1](#), this seems to indicate that both GHMM and HFKM have converged, but the behavior of EM is more difficult to explain.

The jump is explained by a number of “anomalous” trajectories that appear between positions 120 and 160 in the learning data set. These trajectories force the creation of several clusters in the EM model that do not appear in the HFKM approach due to its cluster size threshold. In the case of GHMM, these trajectories also lead to a sudden growth of the model, but this is more difficult to perceive because of its small size.

Finally, [Fig. 51.9](#) plots the model size against the prediction error. This illustrates more explicitly the fact that, for the HFKM and EM algorithms, an increase in the model size does not necessarily lead to an improvement in the accuracy.

5.5 Analysis

Despite the good results so far obtained with GHMMs, more experimental work needs to be done before arriving at definitive conclusions. At this point, results seem to indicate that GHMMs perform better than the similar EM approach mainly because of two factors: (a) they have a more expressive and compact structure and (b) they learn the parameters of the observation and transition probabilities from data instead of relying in a priori knowledge. Even if the first factor is probably more important than the second, it would be

interesting, however, to apply HMM parameter learning to the EM approach after trajectory clustering, to have a quantitative evaluation of the relative importance of both factors. It would also be interesting to do something similar in the case of HFKM, where an HMM is built from the obtained clusters.

6 Conclusions

Most of the approaches discussed here, including all the metric ones, share a major drawback that keeps them of being used in any real-scale application: their reliance in globally localized data renders them unable to generalize the learned knowledge to environments which have not been observed. Moreover, the robot dependence on an instrumented environment makes those algorithms unsuitable for exploration tasks, which are precisely the situations where learning would be most valuable because no other data sources are available to build models by hand.

A better solution would be to use on-board sensors for learning, where the learned behaviors are not anchored to a global frame, but related to specific kinds of places (i.e., roundabouts, crossroads, and parking entrances) that the mobile platform would be able to identify, even in an unknown environment. For such an approach to be possible, it would be necessary to have robust methods for recognizing places by its kind, preferably keeping some sort of semantic soundness. Fortunately, research in this area has seen a considerable boost during the last years thanks to the arrival of new robust 3D range sensors such as the Velodyne, the Swiss ranger, and, more recently, the inexpensive Microsoft Kinect.

In this context, the work of Mozos (2008) is especially relevant, in particular his approach to learning semantic labels of places in range data. The author proposes a supervised learning technique based on AdaBoost that looks promising as the basis for a recognition-based behavior learning approach.

Another mostly unexplored research direction lies in representing object interactions in metric approaches. In principle, there is no fundamental obstacle to integrate techniques already used in intentional approaches, vastly improving prediction soundness and accuracy, and it is likely to see much progress in this area in the short to mid-term.

References

- | | |
|---|--|
| <p>Bennewitz M, Burgard W, Cielniak G, Thrun S (2005) Learning motion patterns of people for compliant robot motion. <i>Int J Robot Res</i> 24(1):31–48</p> <p>Brand M, Oliver N, Pentland A (1997) Coupled hidden markov models for complex action recognition. In: <i>Proceedings of the 1997 conference on computer vision and pattern recognition</i>, San Juan, pp 994–999</p> | <p>Dee H-M (2005) Explaining visible behaviour. PhD thesis, University of Leeds</p> <p>Hoogs A, Perera AA (2008) Video activity recognition in the real world. In: <i>Proceedings of the twenty-third AAAI conference on artificial intelligence</i>, Chicago, pp 1551–1554</p> <p>Hu W, Xiao X, Fu Z, Xie D, Tan T, Maybank S (2006) A system for learning statistical motion patterns.</p> |
|---|--|

- IEEE Trans Pattern Anal Mach Intell 28(9): 1450–1464
- Jockusch J, Ritter H (1999) An instantaneous topological map for correlated stimuli. In: Proceedings of the international joint conference on neural networks, vol 1, Washington DC, pp 529–534
- Juang B-H, Levinson SE, Sondhi MM (1986) Maximum likelihood estimation for multi-variate mixture observations of markov chains. IEEE Trans Inf Theory 32(2):307–309
- Liao L, Fox D, Kautz H (2004) Learning and inferring transportation routines. In: Proceedings of the national conference on artificial intelligence AAAI-04, Amsterdam
- Magee D (2004) Tracking multiple vehicles using foreground, background and shape models. Image Vision Comput 22:143–155
- Makris D, Ellis T (2002) Spatial and probabilistic modelling of pedestrian behavior. In: Proceedings of the British machine vision conference, Cardiff, pp 557–566
- Mozos OM (2008) Semantic place labeling with mobile robots. PhD thesis, University of Freiburg, Freiburg
- Neal RM, Hinton GE (1998) A new view of the em algorithm that justifies incremental, sparse and other variants. In: Jordan MI (ed) Learning in graphical models. Kluwer Academic, Dordrecht, pp 355–368
- Oliver NM, Rosario B, Pentland AE (2000) A Bayesian computer vision system for modeling human interactions. IEEE Trans Pattern Anal Mach Intell 22(8):831–843
- Rabiner LR (1990) A tutorial on hidden markov models and selected applications in speech recognition. Read Speech Recog 77:267–296
- Reif J, Sharir M (1985) Motion planning in the presence of moving obstacles. In Symposium on the foundations of computer science, Portland, pp 144–154
- Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. MIT Press, Cambridge, MA
- Vasquez D (2007) Incremental learning for motion prediction of pedestrians and vehicles. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble. <http://emotion.inrialpes.fr/bibemotion/2007/Vas07>. Accessed 28 Sept 2011
- Vasquez D, Fraichard T, Laugier C (2009) Growing hidden markov models: a tool for incremental learning and prediction of motion. Int J Robot Res 28(11–12):1486–1506
- Walter M, Psarrou A, Gong S (1999) Learning prior and observation augmented density models for behaviour recognition. In: Proceedings of the British machine vision conference, Malvern, Worcestershire, pp 23–32

52 Vision and IMU Data Fusion: Closed-Form Determination of the Absolute Scale, Speed, and Attitude

Agostino Martinelli¹ · Roland Siegwart²

¹INRIA, INRIA Rhone Alpes, avenue de l'Europe, Grenoble, Montbonnot, Saint Ismier, Cedex, France

²Inst.f. Robotik u. Intelligente Systeme, ETHZ, Zurich, Zurich, Switzerland

1	<i>Introduction</i>	1337
2	<i>The Considered System</i>	1339
2.1	The Case with Multiple Features	1340
2.2	The Case with Bias	1341
3	<i>Observability Properties</i>	1341
3.1	The Case Without Gravity	1342
3.2	The Case with Gravity	1343
3.3	The Case with Multiple Features	1344
3.4	The Case with Bias	1344
3.5	Unknown Gravity	1345
4	<i>Necessary Conditions for Observability</i>	1346
5	<i>Closed-Form Solutions</i>	1346
5.1	The Case without Bias	1347
5.2	The Case with Bias	1349

6 *The Algorithm* 1349

7 *Performance Evaluation* 1352

8 *Conclusion* 1353

9 *Appendix A* 1353

9.1 Expression of the Rotation Matrix Ξ by Integrating the Angular Speed1353

Abstract: This chapter describes an algorithm for determining the speed and the attitude of a sensor assembling constituted by a monocular camera and inertial sensors (three orthogonal accelerometers and three orthogonal gyroscopes). The system moves in a 3D unknown environment. The algorithm inputs are the visual and inertial measurements during a very short time interval. The outputs are the speed and attitude, the absolute scale and the bias affecting the inertial measurements. The determination of these outputs is obtained by a simple closed-form solution which analytically expresses the previous physical quantities in terms of the sensor measurements. This closed-form determination allows performing the overall estimation in a very short time interval and without the need of any initialization or prior knowledge. This is a key advantage since allows eliminating the drift on the absolute scale and on the orientation. The performance of the proposed algorithm is evaluated with real experiments.

1 Introduction

In recent years, vision and inertial sensing have received great attention by the mobile robotics community. These sensors require no external infrastructure and this is a key advantage for robots operating in unknown environments where GPS signals are shadowed. In addition, these sensors have very interesting complementarities and together provide rich information to build a system capable of vision-aided inertial navigation and mapping and a great effort has been done very recently in this direction (e.g., Ahrens et al. 2009; Bloesch et al. 2010). A special issue of the *International Journal of Robotics Research* has recently been devoted to the integration of vision and inertial sensors (Dias et al. 2007). In Corke et al. (2007), a tutorial introduction to the vision and inertial sensing is presented. This work provides a biological point of view and it illustrates how vision and inertial sensors have useful complementarities allowing them to cover the respective limitations and deficiencies. The majority of the approaches so far introduced perform the fusion of vision and inertial sensors by filter-based algorithms. In Armesto et al. (2007), these sensors are used to perform egomotion estimation. The sensor fusion is obtained with an Extended Kalman Filter (*EKF*) and with an Unscented Kalman Filter (*UKF*). The approach proposed in Gemeiner et al. (2007) extends the previous one by also estimating the structure of the environment where the motion occurs. In particular, new landmarks are inserted online into the estimated map. This approach has been validated by conducting experiments in a known environment where a ground truth was available. Also, in (Veth and Raquet 2007) an *EKF* has been adopted. In this case, the proposed algorithm estimates a state containing the robot speed, position, and attitude, together with the inertial sensor biases and the location of the features of interest. In the framework of airborne SLAM, an *EKF* has been adopted in Kim and Sukkarieh (2007) to perform 3D-SLAM by fusing inertial and vision measurements. It was remarked that any inconsistent attitude update severely affects any SLAM solution. The authors proposed to separate attitude update from position and velocity update. Alternatively, they proposed

to use additional velocity observations, such as air velocity observation. Regarding the robot attitude, in Bryson and Sukkarieh (2007), it has been noted that roll and pitch angles remain more consistent than the heading.

A fundamental issue to address while fusing vision and inertial measurements is to understand which are the *observable modes*, i.e., the physical quantities that the information contained in the sensor data allows determining. The next issue to address is to find a reliable and efficient method to determine all the previous physical quantities.

The following simple 1-D example clearly shows that it is reasonable to expect that the absolute scale is an observable mode and can be obtained by a closed-form solution. A vehicle equipped with a bearing sensor (e.g., a camera) and an accelerometer moves on a line (see Fig. 52.1). If the initial speed in A is known by integrating the data from the accelerometer, it is possible to determine the robot speed during the subsequent time steps and then the distances $A-B$ and $B-C$ by integrating the speed. The lengths $A-F$ and $B-F$ are obtained by a simple triangulation by using the two angles β_A and β_B from the bearing sensor. When the initial speed v_A is unknown, all the previous segment lengths can be obtained in terms of v_A . In other words, it is possible to obtain the analytical expression of $A-F$ and $B-F$ in terms of the unknown v_A and all the sensor measurements performed while the robot navigates from A to B . By repeating the same computation with the bearing measurements in A and C , it is obtained a further analytical expression for the segment $A-F$ in terms of the unknown v_A and the sensor measurements performed while the vehicle navigates from A to C . The two expressions for $A-F$ provide an equation in the unknown v_A . By solving this equation the value of v_A is obtained. Hence, the value of all the segment lengths in Fig. 52.1 is obtained in terms of the measurements performed by the accelerometer and the bearing sensor.

The previous example is very simple because of several unrealistic restrictions. First of all, the motion is constrained on a line. Additionally, the accelerometer provides gravity-free and unbiased measurements.

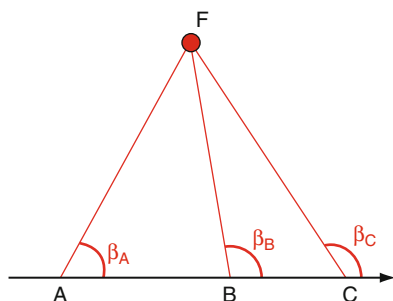


Fig. 52.1

A vehicle equipped with an accelerometer and a camera moves on a line. The camera performs three observations of the feature in F , respectively from the points A , B , and C

In Martinelli (2011b, c), these restrictions were relaxed. A vehicle equipped with IMU and bearing sensors was considered. The motion of the vehicle was not constrained. However, only the case of one single feature was considered. In addition, the inertial measurements were unbiased.

This chapter extends the results obtained in Martinelli (2011b) by also considering the case of multiple features. Additionally, also the case when the accelerometers provide biased measurements will be considered.

The chapter is organized as follows. [Section 2](#) provides a mathematical description of the system. [Sections 3](#) and [4](#) provide conditions for the state observability. Then, [Sect. 5](#) provides the analytical derivation of the closed-form solution to determine the speed and attitude. [Section 6](#) highlights the main steps of the proposed algorithm. The performance of the algorithm is evaluated in [Sect. 7](#). Conclusions are provided in [Sect. 8](#).

2 The Considered System

The system is a sensor assembling constituted by a monocular camera and *IMU* sensors. The IMU consists of three orthogonal accelerometers and three orthogonal gyroscopes. The transformations among the camera frame and the IMU frames are known (the local frame is the camera frame). In the following, the word *vehicle* will be used to refer to this sensor assembling.

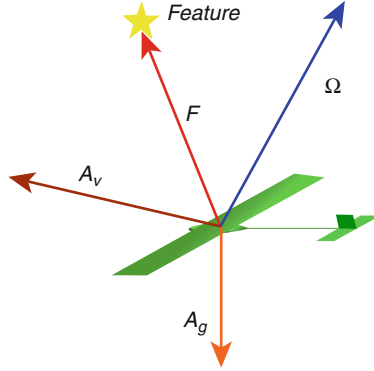
The *IMU* provides the vehicle angular speed and acceleration. Actually, regarding the acceleration, the one perceived by the accelerometer (\mathbf{A}) is not simply the vehicle acceleration (\mathbf{A}_v). It also contains the gravitational acceleration (\mathbf{A}_g). In particular, $\mathbf{A} = \mathbf{A}_v - \mathbf{A}_g$ since, when the camera does not accelerate (i.e., \mathbf{A}_v is zero) the accelerometer perceives an acceleration which is the same of an object accelerated upward in the absence of gravity.

In the following, uppercase letters will indicate the vectors when expressed in the local frame and lowercase letters when they are expressed in the global frame. Hence, regarding the gravity: $\mathbf{a}_g = [0, 0, -g]^T$, where $g \simeq 9.8\text{ms}^{-2}$.

The camera is observing a point feature during a given time interval. The global frame will be attached to this feature. The vehicle and the feature are displayed in [Fig. 52.2](#).

Finally, a quaternion will be adopted to represent the vehicle orientation. Indeed, even if this representation is redundant, it is very powerful since the dynamics can be expressed in a very easy and compact notation (Kuipers 1999).

The system is characterized by the state $[\mathbf{r}, \mathbf{v}, q]^T$, where $\mathbf{r} = [r_x, r_y, r_z]^T$ is the 3D vehicle position, \mathbf{v} is its time derivative, i.e., the vehicle speed in the global frame ($\mathbf{v} \equiv \frac{d\mathbf{r}}{dt}$), $q = q_t + iq_x + jq_y + kq_z$ is a unitary quaternion (i.e., satisfying $q_t^2 + q_x^2 + q_y^2 + q_z^2 = 1$) and characterizes the vehicle orientation. The analytical expression of the dynamics and the camera observations can be easily provided by expressing all the 3D vectors as imaginary quaternions. In practice, given a 3D vector $\mathbf{w} = [w_x, w_y, w_z]^T$,



■ Fig. 52.2

The feature position (F), the vehicle acceleration (A_v), the vehicle angular speed (Ω), and the gravitational acceleration (A_g)

the imaginary quaternion $\hat{w} \equiv 0 + iw_x + jw_y + kw_z$ will be associated with it. The dynamics of the state $[\hat{r}, \hat{v}, q]^T$ are:

$$\begin{cases} \dot{\hat{r}} = \hat{v} \\ \dot{\hat{v}} = q\hat{A}_vq^* = q\hat{A}q^* + \hat{a}_g \\ \dot{q} = \frac{1}{2}q\hat{\Omega} \end{cases} \quad (52.1)$$

where q^* is the conjugate of q , $q^* = q_t - iq_x - jq_y - kq_z$. The camera observations can be expressed in terms of the same state $([\hat{r}, \hat{v}, q]^T)$. The camera provides the direction of the feature in the local frame. In other words, it provides the unit vector $\frac{F}{|F|}$ (see ► Fig. 52.2). Hence, the camera provides the two ratios $y_1 = \frac{F_x}{F_z}$ and $y_2 = \frac{F_y}{F_z}$, where $F = [F_x, F_y, F_z]^T$. The position of the feature in the frame with the same orientation of the global frame but shifted in such a way that its origin coincides with the one of the local frame is $-\hat{r}$. Therefore, F is obtained by the quaternion product $\hat{F} = -q^*\hat{r}q$. The observation function provided by the camera is:

$$h_{cam}(\hat{r}, \hat{v}, q) = [y_1, y_2]^T = \left[\frac{(q^*\hat{r}q)_x}{(q^*\hat{r}q)_z}, \frac{(q^*\hat{r}q)_y}{(q^*\hat{r}q)_z} \right]^T \quad (52.2)$$

where the pedices x , y , and z indicate respectively the i , j , and k components of the corresponding quaternion. Finally, the constraint $q^*q = 1$ can be dealt as a further observation (system output):

$$h_{const}(\hat{r}, \hat{v}, q) = q^*q \quad (52.3)$$

2.1 The Case with Multiple Features

In the case when the camera observes N_f features, simultaneously, the global frame will be attached to one of the features, d_i denotes the 3D vector which contains the cartesian

coordinates of the i^{th} feature ($i = 0, 1, \dots, N_f - 1$). The global frame is attached to the 0^{th} feature, i.e., $\mathbf{d}_0 = [0 \ 0 \ 0]^T$. The new system is characterized by the state $[\hat{r}, \hat{v}, q, \hat{d}_1, \dots, \hat{d}_{N_f-1}]^T$, whose dimension is $7 + 3N_f$. The dynamics of this state are given by (52.1) together with the equation:

$$\dot{\mathbf{d}}_i = [0 \ 0 \ 0]^T \quad i = 1, \dots, N_f - 1 \quad (52.4)$$

The position \mathbf{F}^i of the i^{th} feature in the local frame is obtained by the quaternion product $\hat{\mathbf{F}}^i = q^*(\hat{\mathbf{d}}_i - \hat{r})q$. The corresponding observation function is:

$$h_{cam}^i = \left[\frac{(q^*(\hat{\mathbf{d}}_i - \hat{r})q)_x}{(q^*(\hat{\mathbf{d}}_i - \hat{r})q)_z}, \frac{(q^*(\hat{\mathbf{d}}_i - \hat{r})q)_y}{(q^*(\hat{\mathbf{d}}_i - \hat{r})q)_z} \right]^T, \quad i = 0, 1, \dots, N_f - 1 \quad (52.5)$$

which coincides with the observation in (52.2) when $i = 0$. Summarizing, the case of N_f features is described by the state $[\hat{r}, \hat{v}, q, \hat{d}_1, \dots, \hat{d}_{N_f-1}]^T$, whose dynamics are given in (52.1) and (52.4) and the observations are given in (52.5) and (52.3).

2.2 The Case with Bias

Let us denote with \mathbf{A}_{bias} and $\mathbf{\Omega}_{bias}$ the two 3D vectors whose components are the mean values of the measurement errors from the accelerometers and the gyroscopes, respectively. The two vectors \mathbf{A}_{bias} and $\mathbf{\Omega}_{bias}$ are time-dependent. However, during a short time interval, it is reasonable to consider them to be constant. Under these hypotheses, the dynamics in (52.1) become:

$$\begin{cases} \dot{\hat{r}} &= \hat{v} \\ \dot{\hat{v}} &= q\hat{A}_vq^* = q\hat{A}q^* + q\hat{A}_{bias}q^* + \hat{a}_g \\ \dot{q} &= \frac{1}{2}q\hat{\Omega} + \frac{1}{2}q\hat{\Omega}_{bias} \\ \dot{\mathbf{A}}_{bias} &= \dot{\mathbf{\Omega}}_{bias} = [0 \ 0 \ 0]^T \end{cases} \quad (52.6)$$

Note that these equations only hold for short time intervals. In the following, these equations will be used only when this hypothesis is satisfied (in particular, during time intervals allowing the camera to perform at most ten consecutive observations).

3 Observability Properties

We investigate the observability properties of the system whose dynamics are given in (52.1) and whose observations are given in (52.2 and 52.3). For the sake of clarity, we discuss both the case without gravity (III-A) and with gravity (III-B).

3.1 The Case Without Gravity

Let us set $g=0$ in (52.1). By directly computing the Lie derivatives and their gradients, it is possible to detect three independent symmetries for the resulting system (see Martinelli (2011a)). They are:

$$\begin{aligned} w_s^1 &= \left[0 - r_z r_y 0 - v_z v_y - \frac{q_x q_t}{2} - \frac{q_z q_y}{2} \right]^T \\ w_s^2 &= \left[r_z 0 - r_x v_z 0 - v_x - \frac{q_y q_z q_t}{2} - \frac{q_x}{2} \right]^T \\ w_s^3 &= \left[-r_y r_x 0 - v_y v_x 0 - \frac{q_z}{2} - \frac{q_y q_x q_t}{2} \right]^T \end{aligned} \quad (52.7)$$

According to definition of continuous symmetry introduced in Martinelli (2011a), these vectors are orthogonal to all the gradients of all the Lie derivatives. These symmetries could also be derived by remarking the system invariance with respect to rotations about all the three axes. For instance, an infinitesimal rotation of magnitude ϵ about the vertical axis changes the state as follows (Goldstein 1980):

$$\begin{aligned} \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix} &\rightarrow \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix} + \epsilon \begin{bmatrix} -r_y \\ r_x \\ 0 \end{bmatrix} \\ \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} &\rightarrow \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} + \epsilon \begin{bmatrix} -v_y \\ v_x \\ 0 \end{bmatrix} \\ \begin{bmatrix} q_t \\ q_x \\ q_y \\ q_z \end{bmatrix} &\rightarrow \begin{bmatrix} q_t \\ q_x \\ q_y \\ q_z \end{bmatrix} + \frac{\epsilon}{2} \begin{bmatrix} -q_z \\ -q_y \\ q_x \\ q_t \end{bmatrix} \end{aligned}$$

that is:

$$\begin{bmatrix} r \\ v \\ q \end{bmatrix} \rightarrow \begin{bmatrix} r \\ v \\ q \end{bmatrix} + \epsilon w_s^3$$

On the other hand, without computing the Lie derivatives, we could not conclude that the previous ones are *all* the symmetries for the considered system.

In Martinelli (2011a), we proved that for every symmetry there is an associated partial differential equation and every observable mode must satisfy simultaneously all the three partial differential equations. Since our system is defined by ten variables, the number of independent solutions satisfying all the three partial differential equations is $10-3 = 7$ (John 1982). On the other hand, their derivation, once the three symmetries are detected, is easy. Indeed, it is immediate to prove that the distance of the feature from the camera, i.e., $|r|$, is a solution of the three equations (this can be checked by substitution for the

partial differential equations associated with the symmetries in (52.7) but can also be proved by remarking that the scale factor is invariant under rotations). This means that the distance of the feature is observable and it is one among the seven independent solutions. On the other hand, since the camera provides the position of the feature in the local frame up to a scale factor, having the distance means that the feature position in the local frame is also observable. Therefore, the three components of the feature position in the local frame are three independent solutions. By using quaternions, we can say that three independent solutions are provided by the components of the imaginary quaternion $q^* \hat{r} q$. Additionally, since the three partial differential equations are invariant under the transformation $r \leftrightarrow v$, three other independent solutions are the components of the imaginary quaternion $q^* \hat{v} q$. Physically, this means that the vehicle speed in the local frame is also observable. Finally, the last solution is $q^* q$ since it is directly observed (see 52.3); it can be in any case verified that it satisfies the three partial differential equations.

3.2 The Case with Gravity

We investigate the observability properties when $g \neq 0$. The presence of the gravity breaks two of the previous three symmetries. In other words, the system remains invariant only with respect to rotations about the vertical axis. This means that w_s^1 and w_s^2 are no longer symmetries for the new system. By directly computing the Lie derivatives, we were able to find nine independent Lie derivatives. Hence, the system has $10 - 9 = 1$ symmetry which is w_s^3 .

The partial differential equation associated with w_s^3 is:

$$\begin{aligned} & -2r_y \frac{\partial \Lambda}{\partial r_x} + 2r_x \frac{\partial \Lambda}{\partial r_y} - 2v_y \frac{\partial \Lambda}{\partial v_x} + 2v_x \frac{\partial \Lambda}{\partial v_y} + \\ & -q_z \frac{\partial \Lambda}{\partial q_t} - q_y \frac{\partial \Lambda}{\partial q_x} + q_x \frac{\partial \Lambda}{\partial q_y} + q_t \frac{\partial \Lambda}{\partial q_z} = 0 \end{aligned} \quad (52.8)$$

The number of independent solutions $\Lambda = \Lambda(r_x, r_y, r_z, v_x, v_y, v_z, q_t, q_x, q_y, q_z)$ is equal to the number of variables (i.e., 10) minus the number of equations (i.e., 1) (John 1982). Hence, in this case we have two additional observable modes. They are:

$$Q_r \equiv \frac{q_t q_x + q_y q_z}{1 - 2(q_x^2 + q_y^2)}; \quad Q_p \equiv q_t q_y - q_z q_x \quad (52.9)$$

Also for these two solutions it is possible to find a physical meaning. They are related to the roll and pitch angles (Kuipers 1999). In particular, the first solution provides the roll angle which is $R = \arctan(2Q_r)$. The latter provides the pitch angle which is $P = \arcsin(2Q_p)$. Finally, we remark that the expression of the yaw, $Y = \arctan\left(2 \frac{q_t q_z + q_x q_y}{1 - 2(q_y^2 + q_z^2)}\right)$, does not satisfy (52.8).

3.3 The Case with Multiple Features

Let us suppose that the vehicle is observing $N_f > 1$ features, simultaneously. The new system is characterized by the $(7 + 3N_f)$ – dimensional state $[\hat{r}, \hat{v}, q, \hat{d}_1, \dots, \hat{d}_{N_f-1}]^T$, whose dynamics are given in (52.1) and (52.4) and the observations are given in (52.5 and 52.3).

It is immediate to realize that all the camera observations are invariant with respect to the same symmetries found in the case of one single feature (for instance, the camera observations do not change when the initial state $[\hat{r}, \hat{v}, q, \hat{d}_1, \dots, \hat{d}_{N_f-1}]^T$ is rotated about the vertical axis). Hence, in presence of gravity, the yaw angle is still unobservable. In absence of gravity, also the roll and pitch angles are unobservable. Hence, in presence of gravity, the number of independent modes cannot exceed $7 + 3N_f - 1 = 6 + 3N_f$. In absence of gravity, this number cannot exceed $7 + 3N_f - 3 = 4 + 3N_f$.

On the basis of the results obtained in the previous subsections, we know that the position of each feature in the local frame provides three observable modes. Also, the vehicle speed in the local frame provides three observable modes. In addition, an observable mode is the norm of the quaternion. Therefore, in both the cases with and without gravity, we have $3N_f + 4$ observable modes. In absence of gravity, these are all the observable modes. In presence of gravity, also the roll and pitch angles are observable modes, since they are observable modes with a single feature.

The analytical results derived in the previous subsections can be summarized with the following property:

Property 1

Let us consider the system defined by (52.1, 52.3–52.5). All the independent observable modes are the components of the imaginary quaternion $q^(\hat{d}_i - \hat{r})$, $i = 0, 1, \dots, N_f - 1$ (i.e., the position of the observed features in the local frame), the three components of the imaginary quaternion $q^*\hat{v}q$ (i.e., the vehicle speed in the local frame), and the product q^*q (i.e., the norm of the quaternion). In addition, in presence of gravity, also the roll and pitch angles are observable modes.*

3.4 The Case with Bias

In this subsection we will prove that, even when the camera only observes a single feature, the bias affecting the accelerometers and the gyros are observable. The system we are considering is defined by the state: $[\mathbf{r} \ \mathbf{v} \ q \ \mathbf{b}_A \ \mathbf{b}_\Omega]^T$, whose dimension is 16. This state satisfies the dynamics in (52.6). Finally, this system is characterized by the observations given in (52.2) and (52.3).

We know that the state is not observable. Indeed, even without bias, we know that it is not possible to estimate the yaw angle (Sect. 3.2). In other words, also this system is invariant with respect to rotations about the vertical axis. Hence, its observable modes must satisfy the

equation in (52.8), where, now, Λ also depends on the components of \mathbf{b}_A and \mathbf{b}_Ω . On the other hand, we do not know if the system has additional symmetries in which case the observable modes must satisfy additional partial differential equations, simultaneously. In order to prove that the system has a single symmetry, we must provide 15 independent Lie derivatives. By a direct computation, performed by using the symbolic Matlab computational tool, we were able to find the following 15 independent Lie derivatives: $L^0 y_1, L^0 y_2, L^0 h_{const}, L^1_{f_0} y_1, L^1_{f_0} y_2, L^1_{f_0 f_0} y_1, L^2_{f_0 f_1} y_1, L^2_{f_0 f_4} y_1, L^2_{f_0 f_0} y_2, L^2_{f_0 f_4} y_2, L^2_{f_0 f_5} y_2, L^3_{f_0 f_0 f_5} y_1, L^3_{f_0 f_0 f_6} y_1, L^3_{f_0 f_0 f_2} y_2, L^3_{f_0 f_0 f_6} y_2$. As previously mentioned, we know that we cannot have more than 15 independent Lie derivatives (otherwise, the yaw angle would be observable). The fact that we have 15 independent Lie derivatives means that there are no additional symmetries and, the independent observable modes are the independent solutions of (52.8). They are the nine solutions provided in III-B and the six components of the two vectors \mathbf{b}_A and \mathbf{b}_Ω (note that these components are trivial solutions of (52.8)).

3.5 Unknown Gravity

The results provided in the previous sections are obtained by assuming that the magnitude of the gravity (g) is a priori known. In this section, we want to investigate if the information contained in the sensor data allows us to also estimate g . This investigation could seem useless since in most of the cases the value g is known with good accuracy. On the other hand, this investigation allows us to derive several properties of practical importance.

We will show that g is among the observable modes even in the worst case when the inertial sensors are affected by bias and when only a single feature is available. We will proceed as in Sect. 3.4.

The system we are considering is defined by the state: $[\mathbf{r} \ \mathbf{v} \ \mathbf{q} \ \mathbf{b}_A \ \mathbf{b}_\Omega \ g]^T$, whose dimension is 17. This state satisfies the dynamics in (52.6) with the additional equation $\dot{g} = 0$. Finally, this system is characterized by the observations given in (52.2) and (52.3).

We know that the state is not observable. Indeed, even without bias, we know that it is not possible to estimate the yaw angle (Sect. 3.2). In other words, also this system is invariant with respect to rotations about the vertical axis. Hence, its observable modes must satisfy the equation in (52.8), where Λ also depends on the components of \mathbf{b}_A , \mathbf{b}_Ω and on g . On the other hand, we do not know whether the system has additional symmetries in which case the observable modes must satisfy additional partial differential equations, simultaneously. In order to prove that the system has a single symmetry, we must provide 16 independent Lie derivatives. By a direct computation, performed by using the symbolic Matlab computational tool, we were able to find the following 16 independent Lie derivatives: $L^0 y_1, L^0 y_2, L^0 h_{const}, L^1_{f_0} y_1, L^1_{f_0} y_2, L^2_{f_0 f_0} y_1, L^2_{f_0 f_1} y_1, L^2_{f_0 f_4} y_1, L^2_{f_0 f_0} y_2, L^2_{f_0 f_4} y_2, L^2_{f_0 f_5} y_2, L^3_{f_0 f_0 f_0} y_1, L^3_{f_0 f_0 f_5} y_1, L^3_{f_0 f_0 f_6} y_1, L^3_{f_0 f_3 f_0} y_1, L^3_{f_0 f_0 f_6} y_2$. As previously mentioned, we know that we cannot have more than 16 independent Lie derivatives (otherwise, the yaw angle would be observable). The fact that we have 16

independent Lie derivatives means that there are no additional symmetries and, the independent observable modes are the independent solutions of (52.8). They are the 15 solutions provided in III-D and g .

4 Necessary Conditions for Observability

The observability analysis performed so far takes into account all the degrees of freedom allowed by the dynamics in (52.1). In other words, the observability of the modes previously derived could require the vehicle to move along all these degrees of freedom. It is interesting to understand what happens when only special trajectories are considered. Mathematically, this can be done by introducing in (52.1) the constraints characterizing the trajectory we want to consider. Then, it suffices to apply the method described in Martinelli (2011a) to the system characterized by the new dynamics and the same observations (52.2 and 52.3).

By applying this technique we proved in (Martinelli 2011c) the two following properties:

Property 2

The absolute scale factor is not observable when the vehicle moves at constant speed.

Property 3

When the vehicle moves at constant acceleration all the modes derived in section III are observable, with the exception of the magnitude of the gravitational acceleration (g).

A fundamental consequence of the previous two properties is:

Theorem 1

In order to estimate the observable modes the camera must perform at least three observations (i.e., the observability requires to have at least three images taken from three distinct camera poses). When the magnitude of the gravitational acceleration is unknown, the minimum number of camera images becomes four.

Proof: The proof of this theorem is provided in (Martinelli 2011c). In particular, it is shown that, if the observability of a given physical quantity requires to have a nonconstant speed, this observability also requires at least three camera observations. Similarly, it is shown that, if the observability of a given physical quantity requires to have a nonconstant acceleration, this observability also requires at least four camera observations. ■

In most of the cases, the value g is known with good accuracy. Hence, considering the case of unknown magnitude of gravitational acceleration could seem useless. On the other hand, considering this case has a very practical importance (see Sects. 5 and 6).

5 Closed-Form Solutions

This section provides closed-form solutions which directly express the observable modes in terms of the sensor measurements collected during a short time interval. It starts by dealing with the case without bias.

5.1 The Case without Bias

In the local frame, the dynamics are:

$$\begin{cases} \dot{\mathbf{F}}^i = M\mathbf{F}^i - \mathbf{V} \\ \dot{\mathbf{V}} = M\mathbf{V} + \mathbf{A} + \mathbf{A}_g \\ \dot{\mathbf{q}} = m\mathbf{q} \end{cases} \quad i = 0, 1, \dots, N_f - 1 \quad (52.10)$$

where \mathbf{F}^i is the position of the i^{th} feature in the local frame ($i = 0, 1, \dots, N_f - 1$), \mathbf{V} is the vehicle speed in the same frame, \mathbf{A}_g is the gravitational acceleration in the local frame, i.e., $\hat{\mathbf{A}}_g = \mathbf{q}^* \hat{\mathbf{a}}_g \mathbf{q}$, and \mathbf{q} is the four vector whose components are the components of the quaternion q , i.e., $\mathbf{q} = [q_b, q_x, q_y, q_z]^T$. Finally:

$$m \equiv \frac{1}{2} \begin{bmatrix} 0 & -\Omega_x & -\Omega_y & -\Omega_z \\ \Omega_x & 0 & \Omega_z & -\Omega_y \\ \Omega_y & -\Omega_z & 0 & \Omega_x \\ \Omega_z & \Omega_y & -\Omega_x & 0 \end{bmatrix}, \quad M \equiv \begin{bmatrix} 0 & \Omega_z & -\Omega_y \\ -\Omega_z & 0 & \Omega_x \\ \Omega_y & -\Omega_x & 0 \end{bmatrix}$$

The validity of (52.10) can be checked by using $\hat{\mathbf{F}} = -\mathbf{q}^* \hat{\mathbf{r}} \mathbf{q}$, $\hat{\mathbf{V}} = \mathbf{q}^* \hat{\mathbf{v}} \mathbf{q}$ and by computing their time derivatives with (52.1). In the local frame, the observation in (52.2) for the i^{th} feature is:

$$h_{cam} = [y_1, y_2]^T = \left[\frac{F_x^i}{F_z^i}, \frac{F_y^i}{F_z^i} \right]^T \quad (52.11)$$

Because of the gravity, the first two equations in (52.10) cannot be separated from the equations describing the dynamics of the quaternion.

χ_g will denote the gravity vector in the local frame at a given time T_0 . In other words, $\chi_g \equiv \mathbf{A}_g(T_0)$. Note that, determining χ_g allows determining the roll and pitch angles (R_0 and P_0). Indeed, from the definition of the roll and pitch angles it is possible to obtain:

$$\chi_g = g[\sin P_0, -\sin R_0 \cos P_0, -\cos R_0 \cos P_0]^T \quad (52.12)$$

$\mathbf{F}_0^i \equiv \mathbf{F}^i(T_0)$ will denote the position of the i^{th} feature ($i = 0, 1, \dots, N_f - 1$) at T_0 . Similarly, $\mathbf{V}_0 \equiv \mathbf{V}(T_0)$ will denote the vehicle speed at T_0 .

In the following, a closed-form expression of the vectors $\mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}$, \mathbf{V}_0 , and χ_g in terms of the sensor measurements in the time interval $[T_0, T_0 + T]$ will be provided.

To derive this closed-form expression it is useful to first consider the special case where the vehicle does not rotate during the interval $[T_0, T_0 + T]$. In this case, the first two equations in (52.10) become:

$$\begin{cases} \dot{\mathbf{F}}^i = -\mathbf{V} \\ \dot{\mathbf{V}} = \mathbf{A} + \chi_g \end{cases} \quad i = 0, 1, \dots, N_f - 1 \quad (52.13)$$

It is immediate to integrate the previous equations and obtain the position of the i th feature in the local frame:

$$\mathbf{F}^i(t) = \mathbf{F}_0^i - \Delta t \mathbf{V}_0 - \frac{\Delta t^2}{2} \chi_g - \int_{T_0}^t \int_{T_0}^{t'} \mathbf{A}(\tau) d\tau dt' \quad (52.14)$$

where $\mathbf{A}(\tau)$ are provided by the accelerometers and $\Delta t \equiv t - T_0$.

$\Xi(t)$ will denote the matrix which characterizes the rotation occurred during the interval $[T_0, t]$. The equations in (52.14) correspond to the case when $\Xi(t)$ is the identity 3×3 matrix for any time $t \in [t_0, T_0 + T]$. In the general case, i.e., when the vehicle is not constrained to move with a fixed orientation, $\Xi(t)$ can be evaluated by using the data from the gyroscopes during this time interval (see Appendix A). Hence, it is possible to obtain the extension of (52.14) to a generic motion:

$$\mathbf{F}^i(t) = \Xi(t) \left(\mathbf{F}_0^i - \Delta t \mathbf{V}_0 - \frac{\Delta t^2}{2} \chi_g - \int_{T_0}^t \int_{T_0}^{t'} \Xi^{-1}(\tau) \mathbf{A}(\tau) d\tau dt' \right), \quad i = 0, 1, \dots, N_{f-1} \quad (52.15)$$

In (Martinelli 2011c), the same result has been obtained by directly integrating the equations in (52.10).

The components of $\mathbf{F}^i(t)$, i.e., $F_x^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g)$, $F_y^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g)$ and $F_z^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g)$ are linear in the unknowns $\mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g$. By using (52.11) the following linear equations are obtained:

$$\begin{aligned} F_x^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g) &= \gamma_1(t) F_z^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g) \\ F_y^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g) &= \gamma_2(t) F_z^i(t; \mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g) \end{aligned} \quad (52.16)$$

$i = 0, 1, \dots, N_f - 1$

In particular, each camera observation occurred at the time $t \in [T_0, T_0 + T]$ provides $2N_f$ linear equations in the $3N_f + 6$ unknowns (which are the components of $F_0^i (i = 0, 1, \dots, N_f - 1)$, \mathbf{V}_0 , and χ_g).

When the camera performs observations from n_{obs} distinct poses the number of equations provided by (52.16) is $2n_{obs}N_f$ while the number of unknowns is $3N_f + 6$. In order to determine the unknowns, it is fundamental to know whether these equations are independent or not. To this regard, according to Theorem 1, we know that the number of independent equations is always smaller than the number of unknowns for $n_{obs} \leq 3$. On the other hand, when $n_{obs} = 3$ the knowledge of the magnitude of the gravity makes possible the determination of the modes.

1. $n_{obs} \geq 4$: In this case the number of independent equations in (52.16) is in general larger than the number of unknowns. On the other hand, when $n_{obs} = 4$ and $N_f = 1$, the number of equations is 8, which is less than the number of unknowns 9. In Sect. 6 it is shown that, by using the knowledge of the gravity (i.e., the magnitude of the

vector χ_g), it is possible to determine the unknowns by solving a second order polynomial equation. Hence, in this case, two solutions are determined. When $n_{obs} \geq 5$ and/or $N_f \geq 2$ the determination can be done by the computation of a pseudoinverse. Hence, a single solution can be obtained. Then, the knowledge of the magnitude of the gravitational acceleration can be used to improve the precision (see [Sect. 6](#)).

2. $n_{obs} = 3$: When $N_f = 1$ the number of equations is 6 and the number of unknowns is 9. Hence the estimation cannot be performed. When $N_f \geq 2$ the number of equations is larger or equal to the number of unknowns. On the other hand, according to theorem 1, the vector χ_g cannot be determined, since its norm is not observable. In other words, the equations in ([52.16](#)) are not independent. As in the case $n_{obs} = 4$, $N_f = 1$, it is possible to determine the unknowns by solving a second order polynomial equation. Hence, also in this case, two solutions are determined (see [Sect. 6](#) and (Martinelli 2011c) for further details).

Note that the previous remarks hold in general. There are special situations, whose probability of occurrence is zero, where the determination cannot be carried out. For instance, in the case $n_{obs} = 3$, $N_f = 2$, if one of the three camera poses is aligned along with the two features, the determination cannot be performed. Another special case is when the three camera poses and the two features belong to the same plane.

5.2 The Case with Bias

The closed-form solution will be derived only when the accelerometers are affected by a bias, i.e., in the case $\mathbf{A}_{bias} \neq [0 \ 0 \ 0]^T$ and $\mathbf{\Omega}_{bias} = [0 \ 0 \ 0]^T$.

The expression in ([52.15](#)) can be easily extended to deal with this case by the substitution: $\mathbf{A}(\tau) \rightarrow \mathbf{A}(\tau) + \mathbf{A}_{bias}$:

$$\mathbf{F}^i(t) = \mathbf{\Xi}(t) \left(\mathbf{F}_0^i - \Delta t V_0 - \frac{\Delta t^2}{2} \chi_g - \int_{T_0}^t \int_{T_0}^{t'} \mathbf{\Xi}^{-1} d\tau dt' \mathbf{A}_{bias} - \int_{T_0}^t \int_{T_0}^{t'} \mathbf{\Xi}^{-1}(\tau) \mathbf{A}(\tau) d\tau dt' \right) \quad i = 0, 1, \dots, N_f - 1 \quad (52.17)$$

By proceeding as in the case without bias the analogous of equations ([52.16](#)) is obtained. The new equations also depend on the vector \mathbf{A}_{bias} .

6 The Algorithm

This section describes the algorithm which allows determining the speed, the attitude, and the absolute scale starting from the inertial and visual data collected in a very short time interval. The extension to also determine the bias is straightforward.

As stated at the beginning of the previous section, the local frame is the camera frame. Hence, it is necessary to determine the expression of the acceleration and the angular speed of this frame starting from the acceleration and the angular speed provided by the IMU and from the knowledge of the transformation between the IMU frame and the camera frame (which is assumed a priori known).

The following notation will be adopted:

- C is the matrix which transforms vectors in the IMU frame in vectors expressed in the camera frame
- D is the vector describing the position of the origin of the camera frame in the IMU frame
- Ω_{IMU} and A_{IMU} are the angular speed and the acceleration in the IMU frame (i.e., provided by the inertial sensors)

The expressions of the angular speed and the acceleration in the camera frame are:

$$\Omega = C\Omega_{IMU}, \quad \mathbf{A} = C[A_{IMU} + \dot{\Omega}_{IMU} \wedge \mathbf{D} + \Omega_{IMU} \wedge (\Omega_{IMU} \wedge \mathbf{D})] \quad (52.18)$$

The previous expressions must be used to obtain all the inertial measurements in the camera frame.

The second step consists in computing the matrix Ξ at each time step when the inertial data are delivered (see [Appendix A](#)). This allows computing the term $-\int_{T_0}^t \int_{T_0}^{t'} \Xi^{-1}(\tau) \mathbf{A}(\tau) d\tau dt'$ for all the time t when a camera image is available.

The linear system in [\(52.16\)](#) will be denoted with:

$$\Gamma x = \beta \quad (52.19)$$

where the vector x contains all the unknowns, i.e., $x \equiv [\mathbf{F}_0^0, \mathbf{F}_0^1, \dots, \mathbf{F}_0^{N_f-1}, \mathbf{V}_0, \chi_g]^T$. Γ and β are respectively a $(2n_{obs}N_f \times 3N_f + 6)$ matrix and a $(2n_{obs}N_f \times 1)$ vector and are obtained as follows. For a camera observation occurred at time t , each feature contributes with two rows to the matrix Γ and with two entries to the vector β . For the j^{th} feature observed at time t the three rows of the matrix $\Xi(t)$ will be denoted with $\xi_1(t)$, $\xi_2(t)$, and $\xi_3(t)$. The two rows of the matrix Γ are, respectively:

$$\begin{bmatrix} 0_{3j-3}, \xi_1(t) - y_1(t)\xi_3(t), 0_{3N_f-3j}, -\Delta t(\xi_1(t) - y_1(t)\xi_3(t)), -\frac{\Delta t^2}{2}(\xi_1(t) - y_1(t)\xi_3(t)) \\ 0_{3j-3}, \xi_2(t) - y_2(t)\xi_3(t), 0_{3N_f-3j}, -\Delta t(\xi_2(t) - y_2(t)\xi_3(t)), -\frac{\Delta t^2}{2}(\xi_2(t) - y_2(t)\xi_3(t)) \end{bmatrix}$$

where $\mathbf{0}_n$ denotes the row-vector whose dimension is n and whose entries are all zeros. The corresponding two entries in the vector β are, respectively:

$$[\xi_1(t) - y_1(t)\xi_3(t)] \int_{T_0}^t \int_{T_0}^{t'} \Xi^{-1}(\tau) \mathbf{A}(\tau) d\tau dt', [\xi_2(t) - y_2(t)\xi_3(t)] \int_{T_0}^t \int_{T_0}^{t'} \Xi^{-1}(\tau) \mathbf{A}(\tau) d\tau dt'$$

As stated in the previous section, the matrix Γ is full rank in the following cases:

1. when $n_{obs} \geq 4$ and $N_f \geq 2$
2. when $n_{obs} \geq 5$ and $N_f = 1$

When the rank of Γ is one less than the number of its columns, the nullspace of Γ has dimension one. As discussed at the end of the previous section, this is in general the case when $n_{obs} = 3$, $N_f \geq 2$ (because of the Theorem 1) or when $n_{obs} = 4$, $N_f = 1$. In this case, the system in (52.19) has an infinite number of solutions. By denoting with ν the unit vector belonging to the nullspace of Γ , with x_p one among the solutions of (52.19), any solution of (52.19) is

$$x = x_p + \lambda \nu$$

where λ is a real number. On the other hand, by knowing the magnitude of the gravitational acceleration, it is possible to determine two values of λ . This is obtained by enforcing the constraint that the vector s_λ constituted by the last three entries of the solution $x_p + \lambda \nu$ is a vector with norm equal to g . In other words:

$$|s_\lambda|^2 = g^2 \quad (52.20)$$

which is a second order polynomial equation in λ . Hence, in this case two solutions are determined.

Finally, when Γ is full rank, the knowledge of the magnitude of the gravitational acceleration can be exploited by minimizing the cost function:

$$c(x) = |\Gamma x - \beta|^2 \quad (52.21)$$

under the constraint $|\chi_g| = g$. This minimization problem can be solved by using the method of Lagrange multipliers.

The main steps of the algorithm are displayed in the algorithm 1.

Algorithm 1 (Returns features' positions, speed, and attitude)

Inputs: $\mathbf{A}_{IMU}(t)$, $\Omega_{IMU}(t)$, $y_1^i(t)$, $y_2^i(t)$, ($i = 0, 1, \dots, N_f - 1$), $t \in [T_0, T_0 + T]$

Outputs: \mathbf{F}_0^i , \mathbf{V}_0 , χ_g , ($i = 0, 1, \dots, N_f - 1$)

Compute \mathbf{A} and Ω by using (52.18)

Build the matrix Γ and the vector β in (52.19)

Compute the rank (r) of Γ

if $r = 3N_f + 6$ **then**

$x_{in} = \text{pinv}(\Gamma) \beta$

minimize $c(x)$ in (52.21) with initialization x_{in}

else

if $r = 3N_f + 5$ **then**

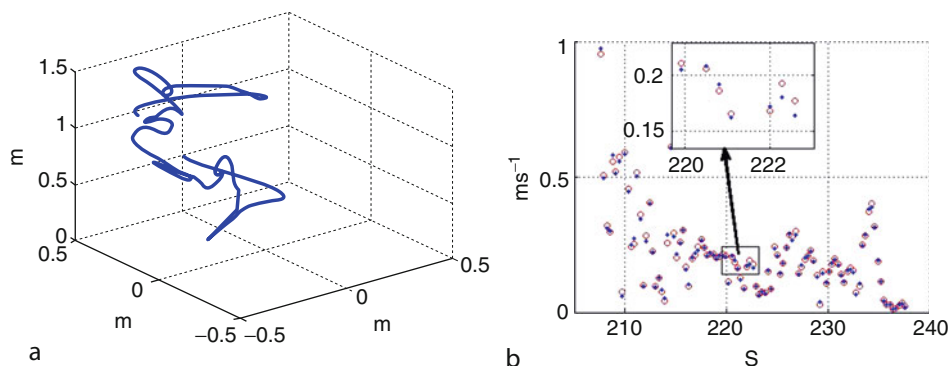
determine two solutions by (52.20)

else

determination not possible

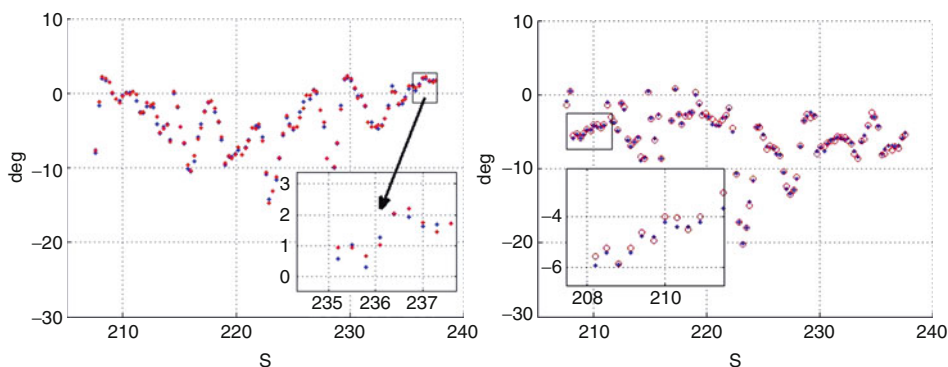
end if

end if



■ Fig. 52.3

In (a) the trajectory (ground truth) in the 3D real data set during the time interval [200, 240]s. In (b) the vehicle speed in the real 3D experiment. *Black dots* are the ground truth and *gray disks* the estimated values



■ Fig. 52.4

Roll (*left*) and pitch (*right*) angles in the real 3D experiment. *Black dots* are the ground truth and *gray disks* the estimated values

7 Performance Evaluation

This section shows the results obtained by using the algorithm with a real data set. The data have been provided by the autonomous system laboratory at ETHZ in Zurich. The data are provided together with a reliable ground truth, which has been obtained by performing the experiments at the ETH Zurich Flying Machine Arena (Lupashin et al. 2010), which is equipped with a Vicon motion capture system. The visual and inertial data are obtained with a monochrome USB-camera gathering 752×480 images at 15 Hz and a Crossbow VG400CC-200 IMU providing the data at 75 Hz. The camera field of view is 150 deg. The calibration of the camera was obtained by using the omnidirectional camera toolkit by Scaramuzza (2006). Finally, the extrinsic calibration between the camera and

the IMU has been obtained by using the strategy introduced in Lobo and Dias (2007). The experiment here analyzed lasted for about 250 s.

► [Figure 52.3a](#) shows the trajectory (ground truth) during the time interval [200, 240]s.

► [Figures 52.3b](#) and ► [52.4](#) show the results regarding the estimated speed, roll, and pitch angles, respectively. In all these figures, the black dots are the ground truth while the gray disks are the estimated values.

8 Conclusion

This chapter describes a method for determining the speed and the attitude of a vehicle equipped with a monocular camera and inertial sensors (i.e., one tri-axial accelerometer and one tri-axial gyrometer). The vehicle moves in a 3D unknown environment. It has been shown that, by collecting the visual and inertial measurements during a very short time interval, it is possible to determine the following physical quantities: the vehicle speed and attitude, the absolute distance of the point features observed by the camera during the considered time interval, and the bias affecting the inertial measurements. In particular, this determination is based on a closed-form solution which analytically expresses the previous physical quantities in terms of the sensor measurements. This closed-form determination allows performing the overall estimation in a very short time interval and without the need of any initialization or a priori knowledge. This is a key advantage since allows eliminating the drift on the scale factor and on the vehicle orientation.

9 Appendix A

9.1 Expression of the Rotation Matrix Ξ by Integrating the Angular Speed

Let us consider a vehicle and a frame attached to this vehicle. When the vehicle moves during the infinitesimal interval $[t_j, t_j + \delta t]$, the rotation matrix which transforms vectors in the reference before this motion and the reference after this motion is: $I_3 + M_j \delta t$, where I_3 is the 3×3 identity matrix and M_j is the skew-symmetric defined in ► [Sect. 5](#) at the time t_j .

Now, let us suppose that the vehicle moves during the interval of time $[t_i, t_f]$. In order to compute the rotation matrix which transforms vectors in the reference before this motion and the reference after this motion, the path is divided in many (N) steps. For each step, the expression of the rotation matrix is the one previously provided. Then, it suffices to compute the product of all these matrices, namely:

$$\Xi = \prod_{k=1}^N (I_3 + M_k \delta t_k) \quad (52.22)$$

where $t_1 = t_i$ and $t_N = t_f$.

References

- Ahrens S, Levine D, Andrews G, How JP (2009) Vision-based guidance and control of a hovering vehicle in unknown, GPS-denied environments. In: IEEE international conference on robotics and automation (ICRA 2009). Kobe, May 2009
- Armesto L, Tornero J, Vincze M (2007) Fast ego-motion estimation with multi-rate fusion of inertial and vision. *Int J Robot Res* 26:577–589
- Blosch M, Weiss S, Scaramuzza D, Siegwart R (2010) Vision based MAV navigation in unknown and unstructured environments. In: IEEE international conference on robotics and automation (ICRA 2010). Anchorage, May 2010
- Bryson M, Sukkariéh S (2007) Building a robust implementation of bearing-only inertial SLAM for a UAV. *J Field Robot* 24:113–143
- Corke P, Lobo J, Dias J (2007) An introduction to inertial and visual sensing. *Int J Robot Res* 26:519–535
- Dias J, Vincze M, Corke P, Lobo J (2007) Editorial: special issue: 2nd workshop on integration of vision and inertial sensors. *Int J Robot Res* 26(6):515–517
- Gemeiner P, Einramhof P, Vincze M (2007) Simultaneous motion and structure estimation by fusion of inertial and vision data. *Int J Robot Res* 26:591–605
- Goldstein H (1980) *Classical mechanics*, 2nd edn. Addison-Wesley, Reading
- John F (1982) *Partial differential equations*. Springer, New York
- Kim J, Sukkariéh S (2007) Real-time implementation of airborne inertial-SLAM. *Robot Auton Syst* 55:62–71
- Kuipers JB (1999) *Quaternions and rotation sequences: a primer with applications to orbits, aerospace, and virtual reality*. Princeton University Press, Princeton copyright
- Lobo J, Dias J (2007) Relative pose calibration between visual and inertial sensors. *Int J Rob Res* 26(6):561–575
- Lupashin S, Schollig A, Sherback M, D'Andrea R (2010) A simple learning strategy for high-speed quadcopter multi-flips. In: IEEE international conference on robotics and automation. Anchorage
- Martinelli A (2011a) State estimation based on the concept of continuous symmetry and observability analysis: the case of calibration. *IEEE Trans Robot* 27(2):239–255
- Martinelli A (2011b) Closed-form solution for attitude and speed determination by fusing monocular vision and inertial sensor measurements, to be presented at the international conference on robotics and automation, ICRA 2011. Shanghai
- Martinelli A (2011c) Vision and IMU data fusion: closed-form solutions for attitude, speed, absolute scale and bias determination. *IEEE Trans on Robot* 28(1)
- Scaramuzza D, Martinelli A, Siegwart R (2006) A toolbox for easy calibrating omnidirectional cameras. In: *Proceedings of the IEEE international conference on intelligent robots and systems (IROS 2006)*. Beijing, October 2006
- Veth M, Raquet J (2007) Fusing low-cost image and inertial sensors for passive navigation. *NAVIGATION* 54(1):11–20

53 Vision-Based Topological Navigation: An Implicit Solution to Loop Closure


Youcef Mezouar^{1,3} · Jonathan Courbon^{1,3} · Philippe Martinet^{2,3}

¹Clermont Université, Université Blaise Pascal, LASMEA

²Clermont Université, IFMA, LASMEA

³CNRS, LASMEA

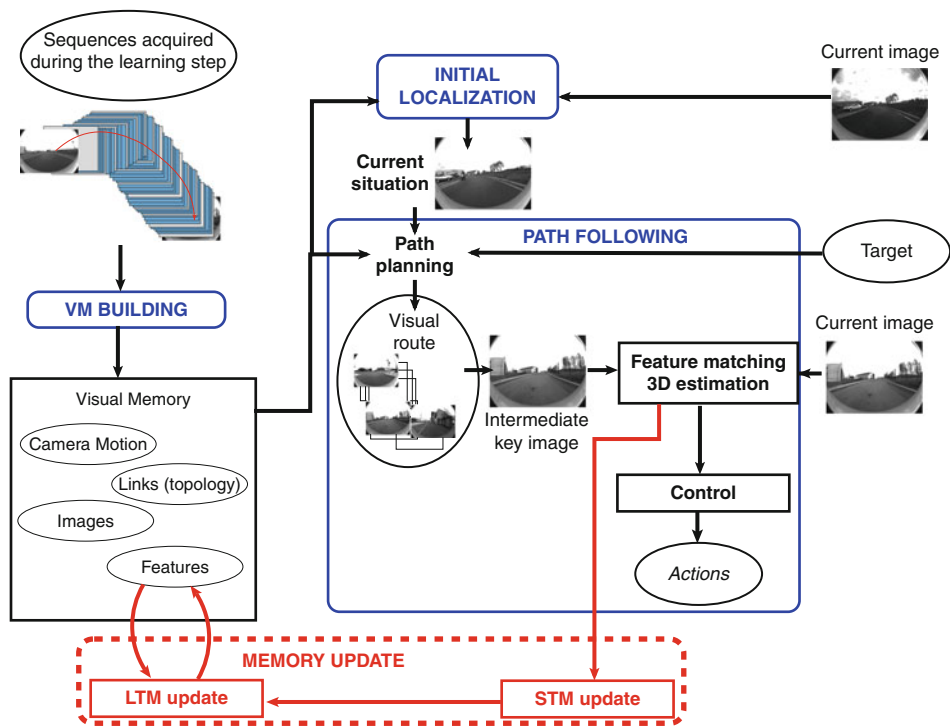
1	Overview	1356
2	Environment Representation	1359
2.1	Visual Memory Structure	1359
2.1.1	Visual Memory	1359
2.1.2	Visual Paths	1359
2.1.3	Visual Memory Vertices	1360
2.1.4	A Connected Multigraph of Weighted Directed Graphs	1361
2.2	Visual Route	1361
2.3	Key Images Selection	1362
2.4	Visual Memory Update	1362
3	Localization in a Visual Memory	1363
3.1	Global Descriptors	1364
3.2	Local Descriptors	1364
3.3	Hybrid Descriptors	1366
4	Route Following	1366
4.1	Model and Assumptions	1367
4.2	Control Design	1370
5	Example of Results	1371
5.1	Experimental Setup	1371
5.2	Loop Closure	1372
5.3	Large Displacement	1373
6	Conclusion	1380

Abstract: Autonomous navigation using a single camera is a challenging and active field of research. Among the different approaches, visual memory-based navigation strategies have gained increasing interests in the last few years. They consist of representing the mobile robot environment with visual features topologically organized gathered in a database (visual memory). Basically, the navigation process from a visual memory can be split in three stages: (1) visual memory acquisition, (2) initial localization, and (3) path planning and following (refer to  [Fig. 53.1](#)). Importantly, this frame work allows accurate autonomous navigation without using explicitly a loop closure strategy. The goal of this chapter is to provide to the reader an illustrative example of such a strategy.

1 Overview

Visual memory-based topological navigation refers to the use of prerecorded and topologically organized 2D image data to drive a robot along a learned trajectory. It relies on techniques inspired from visual servo controls. A major advantage of visual servo control is that absolute geometrical localization of the robot is not required to achieve positioning tasks and thus that drift errors are not propagated along the robot trajectory. However, the use of visual servo control in the field of autonomous navigation faces two major problems: (1) the robot is prone to large displacements which implies that current visual data cannot necessarily be matched with the reference data; (2) conventional visual servo controls make the assumption that a diffeomorphism between the image space and the robot's configuration space exists. Due to the nonholomic constraints of most of wheeled mobile robots, under the condition of rolling without slipping, such a diffeomorphism does not exist if the camera is rigidly fixed to the robot. A potential solution to the first of these two problems is to exploit a suitable environment representation (called visual memory in the sequel) allowing a description of the navigation task as a set of subgoals specified in the observation space. The second problem is often circumvented by providing extra degrees of freedom to the visual sensor. The goal of this chapter is to provide a complete and illustrative framework allowing visual memory-based navigation of non-holonomic wheeled mobile robots without adding extra DoFs to the camera.

The authors of (DeSouza and Kak 2002) account for 20 years of work at the intersection between the robotics and computer vision communities. In many works, as in (Hayet et al. 2002), computer vision techniques are used in a landmark-based framework. Identifying extracted landmarks with known reference points allows to update the results of the localization algorithm. These methods are based on some knowledge about the environment, such as a given 3D model or a map built online. They generally rely on a complete or partial 3D reconstruction of the observed environment through the analysis of data collected from disparate sensors. The vehicle can thus be localized in an absolute reference frame. Both motion planning and vehicle control can then be designed in this space. The results obtained by the authors of (Royer et al. 2007) leave to be forecasted that



■ Fig. 53.1

Navigation process from a visual memory

such a framework will be reachable using a single camera. However, although an accurate global localization is unquestionably useful, the aim of this chapter is to present an alternative to build a complete vision-based framework without recovering the position of the vehicle with respect to a reference frame.

Visual memory-based navigation approaches have gained increasing interest in the last few years. They consist of representing the mobile robot environment with visual features gathered in a database (visual memory). Basically, the navigation process from a visual memory can be split in three stages: (1) visual memory acquisition, (2) initial localization, and (3) path following (refer to ► Fig. 53.1). In the first stage, a sequence of images is acquired, generally during a supervised step, and the robot's internal representation of the environment is built. Basically, three classes of internal representation can be distinguished (DeSouza and Kak 2002): map-less representation, topological and metrical maps. In (Matsumoto et al. 1996), a sequence of images, called *view-sequenced route reference*, is stored in the robot's *brain* for future navigation tasks. Such an approach is ranked among *map-less* as any notion of map or topology of the environment appears, neither to build the reference set of images, nor for the automatic guidance of the mobile robot. More classically, the visual memory is represented by a topological or a metrical

map. In the first case, the nodes of the topological graph represent generally distinctive places while the edges denote connectivity between the places. In metrical maps, the visual memory consists more often of an accurate and consistent 3D representation of the environment. Structure-from-Motion (SfM; Nistér 2004; Royer et al. 2007) and Visual Simultaneous Localization and Mapping (V-SLAM; Lemaire et al. 2007) techniques can be used to build this representation. The SfM problem consists of retrieving the structure of the scene and the motion of the camera using the relation between the views and the correspondences between the features. The number of images of the video sequence initially acquired may be very large and the camera displacement between two views (*baseline*) is however often limited which makes the computation of matching tensors (such as the fundamental matrix) ill conditioned. A solution to decrease this problem is to select a subset of images (*key frames*). Many ways to choose those key images have been proposed (Torr 2002; Pollefeys et al. 2004; Thormählen et al. 2004), balancing the baseline and the number of matched points. Once the key images are chosen, these views, image points, and matched keypoints between successive images can be added to the visual memory. The whole structure of the environment may be built afterward using sequential SfM. Two or three views are usually used to retrieve a first seed 3D structure (Pollefeys et al. 2004; Nistér 2004). Key frames are then sequentially added, computing the pose of each new camera using the previously estimated 3D points (*resection* step). Subsequently, the 3D structure is updated by triangulating the 3D points conveyed by the new view. Both structure and motion are optimized using global (as in Triggs et al. 2000) or local (as in Mouragnon et al. 2009) bundle adjustment. The output of this learning process is a 3D reconstruction of the scene which contains the pose of the camera for each key image and a set of 3D points associated with interest points. The SLAM problem consists of the estimation of the observed environment feature location (*mapping*) and of the robot's pose (*localization*), two problems intimately tied together. Stochastic approaches have proved to solve the SLAM problem in a consistent way because they explicitly deal with sensor noise. A feature-based SLAM approach generally encompasses four basic functionalities: feature selection, relative measures estimation, data association, and estimation. In V-SLAM, the observed features can be for instance interest points detected in the images and data association performed by a feature matching process. Filters like the Extended Kalman Filter are then used to estimate both the localization of the robot and the 3D position of features on the environment. The second stage of the navigation process (initial localization) consists of finding the position of the robot in its internal representation of the environment using the current image acquired by the embedded camera. It can rely on image matching and/or on the matching of features extracted from the current image and images stored in the visual memory. Once the robot is localized and a target is specified in its internal representation of the environment, the next stage (navigation) consists first in planning the robot's mission and second to perform it autonomously. In the sequel, this chapter will focus on navigation strategies where key images are stored in the visual memory and are used as references during the online steps.

2 Environment Representation

In (DeSouza and Kak 2002), approaches using a “memorization” of images of the environment acquired with an embedded camera are ranked among map-less navigation systems. As proposed in (Matsumoto et al. 1996) or (Jones et al. 1997), neither notion of mapping nor topology of the environment appears, in building the reference set of images, nor for the automatic guidance of the vehicle. The first step in vision-based topological navigation strategies consists of a learning stage to build the visual memory.

2.1 Visual Memory Structure

The environment is supposed to contain a set of 3D features $\{Q_l \mid l = 1, 2, \dots, n\}$. The observation (or projection) of a 3D feature Q_l in an image \mathcal{I}^{i_a} is a visual feature noted \mathcal{P}_l^* (refer to ⑤ Fig. 53.2). It is assumed that visual features can be located/detected from images and that they are described by feature vectors. Two features $\mathcal{P}_{l_1}^{i_1}$ and $\mathcal{P}_{l_2}^{i_2}$ from two images \mathcal{I}^{i_1} and \mathcal{I}^{i_2} are said to be *matched* or *in correspondence* if they are supposed to be the projections of a same 3D feature (i.e., $l_1 = l_2$).

2.1.1 Visual Memory

The visual memory of the robot can store different features. In this chapter, the concept of visual memory is illustrated assuming that the following 2D features are stored:

- (a) n_{VM} key images $\{\mathcal{I}^i \mid i = \{1, 2, \dots, n_{VM}\}\}$ extracted from a video sequence
- (b) For each key image \mathcal{I}^i , a set P^i of n^i descriptive image features

$$P^i = \left\{ \mathcal{P}_{l_j}^i \mid j = \{1, 2, \dots, n^i\}, l_j \in \{1, 2, \dots, n\} \right\}$$
- (c) A set of links between adjacent places $\{(\mathcal{I}^{i_a}, \mathcal{I}^{i_b}), (i_a, i_b) \in \{1, 2, \dots, n_{VM}\}^2, i_a \neq i_b\}$

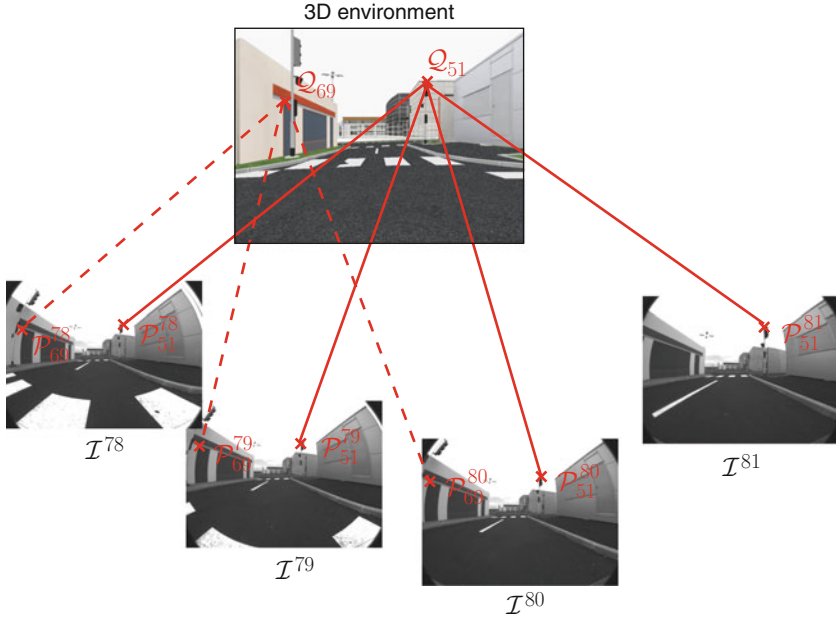
2.1.2 Visual Paths

A visual path Ψ^P is a weighted directed graph composed of n successive key images (*vertices*):

$$\Psi^P = \{\mathcal{I}_i^P \mid i \in \{1, 2, \dots, n\}\}$$

For control purpose (refer to ⑤ Sect. 4), the authorized motions during the learning stage are assumed to be limited to those of a car-like vehicle, which only goes forward. The following Hypothesis 1 formalizes these constraints.

Hypothesis 1: Given two frames ${}^R\mathcal{F}_i$ and ${}^R\mathcal{F}_{i+1}$, respectively associated to the vehicle when two successive key images \mathcal{I}_i and \mathcal{I}_{i+1} of a visual path Ψ were acquired, there exists an



■ Fig. 53.2

Images, 3D features, and visual features

admissible path ψ from ${}^R\mathcal{F}_i$ to ${}^R\mathcal{F}_{i+1}$ for a car-like vehicle whose turn radius is bounded, and which only moves forward.

Moreover, because the controller is assumed vision based, the vehicle is controllable from I_i to I_{i+1} only if the hereunder Hypothesis 2 is respected.

Hypothesis 2: Two successive key images I_i and I_{i+1} contain a set P_i of matched visual features, which can be observed along a path performed between ${}^R\mathcal{F}_i$ and ${}^R\mathcal{F}_{i+1}$ and which allows the computation of the control law.

In the sequel, this chapter is illustrated using interest points as visual features. During the acquisition of a visual path, the Hypothesis 2 constrains the choice of the key images. As a consequence of Hypothesis 1 and 2, each visual path Ψ^p corresponds to an oriented edge which connects two configurations of the vehicle's workspace. The *weight of a visual path* can be defined for instance as its cardinal.

2.1.3 Visual Memory Vertices

In order to connect two visual paths, the terminal extremity of one of them and the initial extremity of the other one must be constrained as two consecutive key images of a visual path. The paths are then connected by a vertex, and two adjacent vertices of the visual memory are connected by a visual path.

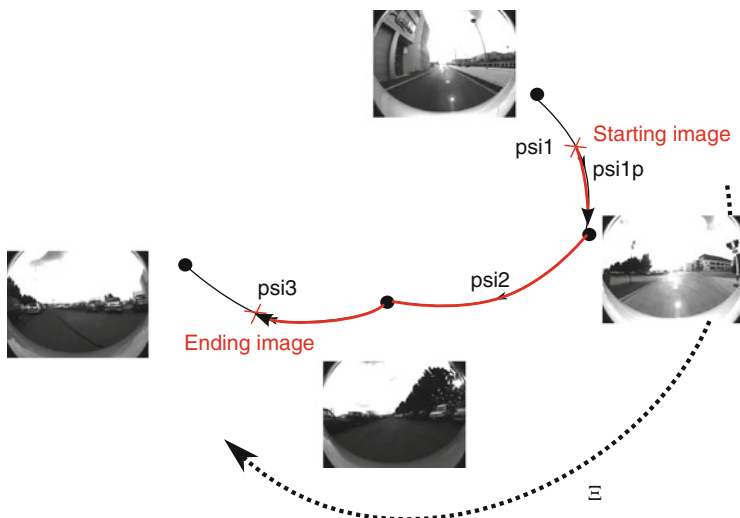
Proposition 1: Given two visual paths $\Psi^{p_1} = \{\mathcal{I}_i^{p_1} | i \in \{1, 2, \dots, n_1\}\}$ and $\Psi^{p_2} = \{\mathcal{I}_i^{p_2} | i \in \{1, 2, \dots, n_2\}\}$, if the two key images $\mathcal{I}_{n_1}^{p_1}$ and $\mathcal{I}_1^{p_2}$ abide by both Hypothesis 1 and 2, then a vertex connects Ψ^{p_1} to Ψ^{p_2} .

2.1.4 A Connected Multigraph of Weighted Directed Graphs

According to ▶ Sects. 2.1.2 and ▶ 2.1.3, the visual memory structure is a multigraph in which vertices are key images linked by edges which are the visual paths (*directed graphs*). Note that more than one visual path may be incident to a node. It is yet necessary that this multigraph is strongly connected. This condition guarantees that any vertex of the visual memory is attainable from every other, through a set of visual path.

2.2 Visual Route

A visual route describes the vehicle's mission in the sensor space. Given two key images of the visual memory \mathcal{I}_s^* and \mathcal{I}_g , corresponding respectively to the starting and goal locations of the vehicle in the memory, a visual route is a set of key images which describes a path from \mathcal{I}_s^* to \mathcal{I}_g , as presented in ▶ Fig. 53.3. \mathcal{I}_s^* is the closest key image to the current image \mathcal{I}_s . The image \mathcal{I}_s^* is extracted from the visual memory during a localization step. The visual route can be chosen for instance as the minimum length path of the visual memory connecting two vertices associated to \mathcal{I}_s^* and \mathcal{I}_g . According to the definition of the



■ Fig. 53.3

The tasks consists of navigating from the starting to the ending images. With this aim, a visual route $\Xi = \Psi^{1'} \oplus \Psi^2 \oplus \Psi^{3'}$ connecting these two images is defined

value of a visual path, the length of a path is the sum of the values of its arcs. Consequently, the visual route results from the concatenation of indexed visual paths. Given two visual paths Ψ^{p_1} and Ψ^{p_2} , respectively containing n_1 and n_2 indexed key images, the concatenation operation of Ψ^{p_1} and Ψ^{p_2} is defined as follows:

$$\Psi^{p_1} \oplus \Psi^{p_2} = \left\{ \mathcal{I}_j^{p_{1,2}} \mid j = \{1, \dots, n_1, n_1 + 1, \dots, n_1 + n_2\} \right\}$$

$$\mathcal{I}_j^{p_{1,2}} = \begin{cases} \mathcal{I}_j^{p_1} & \text{if } j \leq n_1 \\ \mathcal{I}_{j-n_1}^{p_2} & \text{if } n_1 < j \leq n_1 + n_2 \end{cases}$$

2.3 Key Images Selection

A central clue for implementation of this framework relies on efficient point matching. It allows key image selection during the learning stage, of course it is also useful during autonomous navigation in order to provide the necessary input for state estimation. A simple but efficient solution to this issue is given in (Royer et al. 2007) and was successfully applied for the metric localization of autonomous vehicles in outdoor environment. Interest points are detected in each image with Harris corner detector (Harris and Stephens 1988). For an interest point \mathcal{P}_1 at coordinates $(x \ y)$ in image I_i , a search region in image I_{i+1} is defined. For each interest point \mathcal{P}_2 inside the search region in image I_{i+1} , a similarity score is computed between the neighborhoods of \mathcal{P}_1 and \mathcal{P}_2 using a zero-normalized cross correlation. The point with the best score is kept as a good match and the unicity constraint is used to reject matches which have become impossible. This method is illumination invariant and its computational cost is small. The first image of the video sequence is selected as the first key frame I_1 . A key frame I_{i+1} is then chosen so that there are as many video frames as possible between I_i and I_{i+1} while there are at least M common interest points tracked between I_i and I_{i+1} .

2.4 Visual Memory Update

The internal representation of the environment is generally built once and never changed. Most navigation strategies proposed in the literature assume that the environment where the robot works is static. However, this assumption does not hold for many real environments. Following the taxonomy proposed in (Yamauchi and Langley 1997), changes in the environment may be *transient* or *lasting*. Transient changes are brief enough and can be handled reactively. In general, it does not require any long-standing modification of the robot's internal memory (for instance, moving objects or walking pedestrians). Lasting changes persist over longer periods of time and have to be memorized by the robot. They may be *topological* (changes in the topology) and/or *perceptual* (changes in the appearance of the environment).

As noted previously, perceptual lasting changes will deteriorate the feature-matching process and then the performance of vision-based navigation strategies. To improve the navigation performances, new lasting features have to be incorporated in the map of the environment. Further, obsolete elements have to be eliminated to limit the required resources in terms of memory and processing power over time.

As mentioned previously, a large part of the literature deals with transient changes. The robot's environment is generally decomposed into a static part and a dynamic part encapsulating ephemeral (potentially moving) objects. Two solutions can be used to deal with this situation. The first solution consists of identifying the parts of the environment which are not consistent with a predefined static model. This is usually bypassed with geometric consistency of view matching. The second solution consists of tracking moving objects as proposed in the context of V-SLAM in (Bibby and Reid 2007; Wangsiripitak and Murray 2009). These objects can then be integrated to the map building process as in (Bibby and Reid 2007) or rejected as in (Wangsiripitak and Murray 2009). However, these solutions may improve the current localization but cannot handle long-term changes on the structure of the environment.

Only few works have been devoted to lasting changes. In feature-based visual SLAM approaches, features accumulate over time (which can be seen as a map update) but obsolete features are not discarded. It results a growing of the required memory and processing power over time and an efficiency loss. In (Hochdorfer and Schlegel 2009), the evaluation of the quality of the localization allows to rank landmarks and to eliminate less useful ones. In (Andreasson et al. 2007), the initial map is supposed to be partially correct and a robust method for global place recognition in scenes subject to changes over long periods of time is proposed. As the reference view is never modified, this approach may be inefficient after some times. It seems more promising to modify the reference views as proposed in (Dayoub and Duckett 2008; Dayoub et al. 2010; Bacca et al. 2010) for localization. The information model used in those works is based on the human memory model proposed in (Atkinson and Shiffrin 1968) and the concepts of short-term and long-term memories. Basically, reference views are stored in a long-term memory (LTM). When features have been seen in many views during the localization step, they are transferred from the short-term memory (STM) to the long-term memory (if they do not belong yet to it) and missing features are forgotten (and are deleted after sometime). The updates of the memories are based on a finite state machine in (Dayoub and Duckett 2008; Dayoub et al. 2010) and on feature stability histograms built using a voting scheme in (Bacca et al. 2010). Those approaches are tested with images acquired by omnidirectional cameras in indoor environments. It is reported that localization performances are improved with respect to a static map.

3 Localization in a Visual Memory

The output of the learning process is a data set of images (*visual memory*). The first step in the autonomous navigation process is the self-localization of the vehicle in the visual memory. In a visual memory, the localization consists of finding the image of the memory

which best fits the current image by comparing preprocessed and online acquired images. Two main strategies exist to match images: The image can be represented by a single descriptor (global approaches) (Matsumoto et al. 1999; Linåker Fand and Ishikawa 2004) or alternatively by a set of descriptors defined around visual features (landmark-based or local approaches) (Goedemé et al. 2005; Tamimi et al. 2005; Murillo et al. 2007). Some hybrid approaches based on a global description of a subset of the image have also been proposed to increase the robustness of global methods (Gonzalez-Barbosa and Lacroix 2002). On the one hand, local approaches are generally more accurate but have a high computational cost (Murillo et al. 2007). On the other hand, global descriptors speed up the matching process at the price of affecting the robustness to occlusions. One solution consists in using a hierarchical approach which combines the advantages of both methods (Menegatti et al. 2003). In a first step, global descriptors allow to select only some possible images and then, if necessary, local descriptors are used to keep the best image. This section briefly reviews global and local descriptors for localization in a visual memory with a particular focus on wide field-of-view images since they are of particular interests in the context of autonomous navigation.

3.1 Global Descriptors

A first solution is to globally describe the image. In that aim, images are mapped onto cylindrical images of size 128×32 in (Matsumoto et al. 1999). The image is directly described by the gray-level values. In (Pajdla Tand and Hlaváč 1999), a shift invariant representation is computed by rotating the cylindrical image in a reference direction. Unfortunately, this direction is not absolute as soon as occlusions appear. In order to decrease the size of the memorized data, images can be represented by their eigenvectors using principal component analysis as proposed in (Gaspar et al. 2000). Unfortunately, when a new image is integrated in the memory, all eigenvectors have to be recomputed. This process is very complex and it has a very high computational cost. Moreover, those methods are not robust to changes of the environment. The histogram of the gray-level values is largely employed as global signature. Its computation is efficient and it is rotation-invariant. However, histogram methods are sensitive to change of light conditions. Blaer and Allen (2002) propose color histograms for outdoor scene localization. A normalization process is applied before computing the histograms in order to reduce the illumination variations. In (Linåker Fand and Ishikawa 2004), a global descriptor based on a polar version of high order local autocorrelation functions (PHLAC) is proposed. It is based on a set of 35 local masks applied to the image by convolution. Similar to histogram, this descriptor is rotation-invariant.

3.2 Local Descriptors

Global descriptor-based methods are generally less robust to occlusion compared to landmark-based methods. In those last methods, some relevant visual features are

extracted from the images. A descriptor is then associated to each feature neighborhood. The robustness of the extraction and the invariance of the descriptor are one main issue to improve the matching process. Two main approaches can be distinguished. In the first category, the feature detection and description designed for images acquired by perspective cameras are directly employed with omnidirectional images. The second category takes the geometry of the sensor into account and thus uses operators designed for omnidirectional images. The most popular visual features used in the context of localization in an image database are projected points. However, projected lines can also be exploited as proposed in (Murillo et al. 2007).

1. *Perspective-based local descriptor*: The Scale Invariant Feature Transform (SIFT, (Lowe 2004)) has been shown to give the best results in the case of images acquired with perspective cameras. The SIFT descriptor is a set of histograms of gradient orientations of the normalized (with respect to orientation and scale) difference of Gaussian images. In view of the effectiveness of this descriptor, several extensions have been proposed. It has been used with omnidirectional images in (Goedemé et al. 2005). Given that many points are detected in an omnidirectional image, Tamimi et al. (2005) proposed an iterative SIFT with a lower computational cost. In (Andreasson et al. 2005), points are detected with a Sobel filter and described by a Modified Scale Invariant Feature Transform (M-SIFT) signature. This signature slightly takes into account the sensor geometry by rotating the patch around an interest point. In (Murillo et al. 2007), the Speeded-Up Robust Features (SURF) are employed as descriptors. SURF points are detected using the Hessian matrix of the image convolved with box filters and the descriptor is computed thanks to Haar wavelet extraction. The computational cost of this descriptor is much lower than the one obtained for SIFT. Unfortunately, those signatures describe a local neighborhood around interest points and do not take into account the high distortions caused by the sensor geometry.
2. *Descriptors adapted to wide-angle images*: In the second category, detection and description processes are specially designed to take into account high distortions. In (Svoboda and Pajdla 2001; Ieng et al. 2003), a classical Harris corner detector is proposed but the shape and the size of a patch around a feature is modified according to the position of the point and to the geometry of the catadioptric sensor. Finally, a standard 2D correlation (respectively a centered and normalized cross correlation) is applied to the patches in (Svoboda and Pajdla 2001) (respectively in Ieng et al. 2003). After computing the descriptors of the current and memorized images, those descriptors have to be matched. For local approaches, this step is generally based on pyramidal matching as in (Murillo et al. 2007) or on nearest neighbor matching as in (Lowe 2004). This last algorithm considers that a matching is correct if the ratio between the distances of the first and second nearest neighbors is below a threshold. It is possible to eliminate wrong matching through the recovery of the epipolar geometry between two views (Zhang et al. 1995) at the price of higher computational cost. A full reconstruction can also be obtained with three views and the 1D trifocal tensor as proposed in (Murillo et al. 2007).


3.3 Hybrid Descriptors

Some hybrid descriptors have been designed to combine the advantages of the two previously cited categories (local and global approaches) by globally describing subsets of the image. In (Gonzalez-Barbosa and Lacroix 2002), five histograms of the first- and second-order derivatives of the gray-level image are considered. Instead of the whole image, the image is decomposed into rings. On the one hand, a decomposition into few rings decreases the accuracy. On the other hand, increasing the number of rings increases the computational cost and decreases the robustness to occlusions. In (Gaspar et al. 2000), the image is first projected onto an englobing cylinder and a grid decomposition is then proposed. This projection step is time consuming and it implies the modification of the quality of the image which can lead to less accurate localization results. In (Courbon et al. 2008), a hierarchical process combining global descriptors computed onto cubic interpolation of triangular mesh and patches correlation around Harris corners has been proposed. In the context of visual memory-based navigation, this method has shown the best compromise in terms of accuracy, amount of memorized data required per image, and computational cost (refer to (Courbon et al. 2008) for detailed results).



4 Route Following

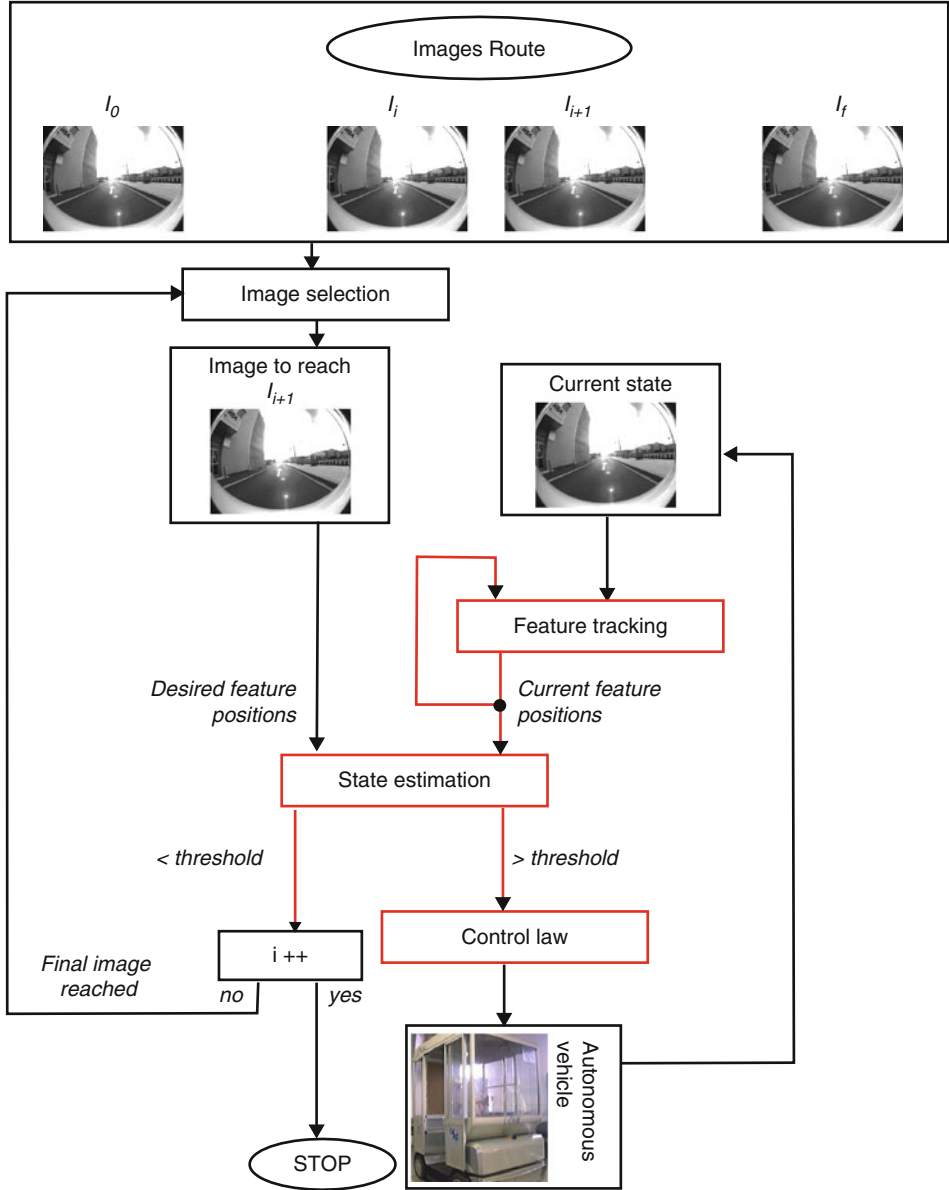
Given an image of one of the visual paths as a target, the navigation task in a visual memory-based framework can formally be defined as the regulation of successive error functions allowing the guidance of the robot along the reference visual route. The visual route describes then a set of consecutive states that the image has to reach in order that the robot joins the goal configuration from the initial one. Control schemes suitable in this context can be designed by exploiting visual-servoing concepts. Visual servoing is often considered as a way to achieve positioning tasks. Classical methods, based on the task function formalism, make the assumption that a diffeomorphism between the sensor space and the robot's configuration space exists. Due to the nonholomic constraints of most of wheeled mobile robots, under the condition of rolling without slipping, such a diffeomorphism does not exist if the camera is rigidly fixed to the robot. In (Tsakiris et al. 1998), the authors add extra degrees of freedom to the camera. The camera pose can then be regulated in a closed loop. In the case of an embedded and fixed camera, the control of the camera is generally based on wheeled mobile robots control theory (Samson 1995). In (Ma et al. 1999), a car-like robot is controlled with respect to the projection of a ground curve in the image plane. The control law is formalized as a path-following problem. More recently, in (Fang et al. 2002) and (Chen et al. 2003), a partial estimation of the camera displacement between the current and desired views has been exploited to design vision-based control laws. The camera displacement is estimated by uncoupling translation and rotation components of an homography matrix. In (Fang et al. 2002), a time-varying control allows an asymptotical

stabilization on a desired image. In (Chen et al. 2003), a trajectory-following task is achieved. The trajectory to follow is defined by a prerecorded video and the control law is proved stable using Lyapunov-based analysis. In (Goedemé et al. 2005), homing strategy is used to control a wheelchair from a memory of omnidirectional images. A memory of omnidirectional images is also used in (Gaspar et al. 2000) where localization and navigation are realized in the bird's-eye (orthographic) views obtained by radial distortion correction of the omnidirectional images. The control of the robot is formulated in the bird's-eye view of the ground plane which is similar to a navigation in a metric map. The view-sequenced route presented in (Matsumoto et al. 1996) has been applied to omnidirectional images in (Matsumoto et al. 1999). The control scheme exploits the inputs extracted from unwarped images. For completeness, the control strategy proposed in (Courbon et al. 2009) to follow a visual route with a non-holonomic vehicle is briefly presented more in details.

The localization step provides the closest image \mathcal{I}_s^* to the current initial image I_c . A visual route Ψ connecting \mathcal{I}_s^* to the goal image can then be extracted from the visual memory. The principle of the vision-based control scheme is presented in  Fig. 53.4.

4.1 Model and Assumptions

1. *Control objective:* Let I_i and I_{i+1} be two consecutive key images of a given visual route to follow and I_c be the current image. $\mathcal{F}_i = (O_i, \mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)$ and $\mathcal{F}_{i+1} = (O_{i+1}, \mathbf{X}_{i+1}, \mathbf{Y}_{i+1}, \mathbf{Z}_{i+1})$ are the frames attached to the vehicle when I_i and I_{i+1} were stored and $\mathcal{F}_c = (O_c, \mathbf{X}_c, \mathbf{Y}_c, \mathbf{Z}_c)$ is a frame attached to the vehicle in its current location.  Figure 53.5 illustrates this setup. The origin O_c of \mathcal{F}_c is on the center rear axle of a car-like vehicle, which moves on a perfect ground plane. The hand-eye parameters (i.e., the rigid transformation between \mathcal{F}_c and the frame attached to the camera) are supposed to be known. According to Hypothesis 2, the state of a set of visual features \mathcal{P}_i is known in the images I_i and I_{i+1} . The state of \mathcal{P}_i is also assumed available in I_c (i.e., \mathcal{P}_i is in the camera field of view). The task to achieve is to drive the state of \mathcal{P}_i from its current value to its value in I_{i+1} . In the following, Γ represents a path from \mathcal{F}_i to \mathcal{F}_{i+1} . The control strategy consists in guiding I_c to I_{i+1} by regulating asymptotically the axle \mathbf{Y}_c on Γ . The control objective is achieved if \mathbf{Y}_c is regulated to Γ before the origin of \mathcal{F}_c reaches the origin of \mathcal{F}_{i+1} .
2. *Vehicle modeling:* The vehicle is supposed to move on asphalt at rather slow speed. In this context, it appears quite natural to rely on a kinematic model, and to assume pure rolling and nonslipping at wheel-ground contact. In such cases, the vehicle modeling is commonly achieved for instance relying on the Ackermann's model, also named the bicycle model: the two front wheels located at the mid-distance between actual front wheels and actual rear wheels. In the sequel, the robot configuration is described with respect to the path Γ , rather than with respect to an absolute frame. As seen previously, the objective is that the vehicle follows a reference path Γ . To meet this objective, the following notations are introduced (see  Fig. 53.5).



■ Fig. 53.4
Visual route following process

- O_C is the center of the vehicle rear axle.
- \mathcal{M} is the point of Γ which is the closest to O_C . This point is assumed to be unique which is realistic when the vehicle remains close from Γ .
- s is the curvilinear coordinate of point M along Γ and $c(s)$ denotes the curvature of Γ at that point.

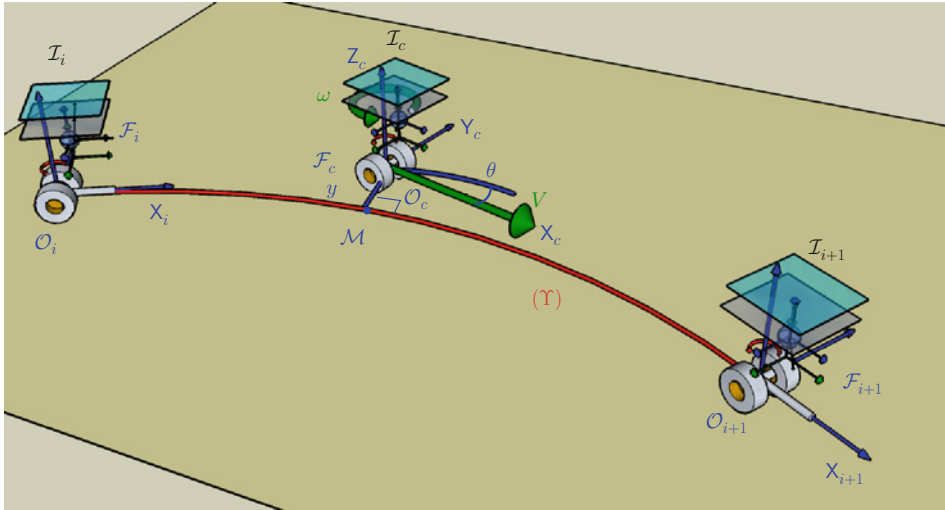


Fig. 53.5

Images \mathcal{I}_i and \mathcal{I}_{i+1} are two consecutive key images of the visual route Ψ . \mathcal{I}_c is the current image. Γ is the path to follow

- y and θ are respectively the lateral and angular deviation of the vehicle with respect to reference path Γ .
- δ is the virtual front wheel steering angle.
- V is the linear velocity along the axle \mathbf{Y}_c of \mathcal{F}_c .
- l is the vehicle wheelbase.

Vehicle configuration can be described without ambiguity by the state vector (s, y, θ) : The two first variables provide point O_C location and the last one the vehicle heading. Since V is considered as a parameter, the only control variable available to achieve path following is δ . The vehicle kinematic model can then be derived by writing that velocity vectors at point O_C and at center of the front wheel are directed along wheel planes and that the vehicle motion is, at each instant, a rotation around an instantaneous rotation center. Such calculations lead to (refer to Zodiac 1995):

$$\begin{cases} \dot{s} = V \frac{\cos \theta}{1 - c(s)y} \\ \dot{y} = V \sin \theta \\ \dot{\theta} = V \left(\frac{\tan \delta}{l} - \frac{c(s) \cos \theta}{1 - c(s)y} \right) \end{cases} \quad (53.1)$$

Model (53.1) is clearly singular when $y = \frac{1}{c(s)}$, i.e., when point O_C is superposed with the path Γ curvature center at abscissa s . However, this configuration is never encountered in practical situations: On the one hand, the path curvature is small and on the other, the vehicle is expected to remain close to Γ .

4.2 Control Design

The control objective is to ensure the convergence of y and θ toward 0 before the origin of \mathcal{F}_c reaches the origin of \mathcal{F}_{i+1} . The vehicle model (► 53.1) is clearly nonlinear. However, it has been established in (Samson 1995) that mobile robot models can generally be converted in an exact way into almost linear models, named chained forms. This property offers two very attractive features: On the one hand, path following control law can be designed and tuned according to Linear System Theory, while controlling nevertheless the actual nonlinear vehicle model. Control law convergence and performances are then guaranteed whatever the vehicle initial configuration is. On the other hand, chained form enables to specify, in a very natural way, control law in terms of distance covered by the vehicle, rather than in terms of time. Vehicle spacial trajectories can then easily be controlled, whatever the vehicle velocity is (Thuilot et al. 2004). Conversion of the vehicle model (► 53.1) into chained form can be achieved thanks to state and control transformations as detailed in (Thuilot et al. 2004) leading to the following expression of the control law:

$$\delta(y, \theta) = \arctan \left(-l \left[\frac{\cos^3 \theta}{(1 - c(s)y)^2} \left(\frac{dc(s)}{ds} y \tan \theta - K_d(1 - c(s)y) \tan \theta - K_p y + c(s)(1 - c(s)y) \tan^2 \theta \right) + \frac{c(s) \cos \theta}{1 - c(s)y} \right] \right) \quad (53.2)$$


The evolution of the error dynamics is driven by the distance covered by the vehicle along the reference path Γ). The gains (K_d , K_p) impose a settling distance instead of a settling time as it is usual. Consequently, for a given initial error, the vehicle trajectory will be identical, whatever the value of V is, and even if V is time varying ($V \neq 0$). Control law performances are therefore velocity independent. The gains (K_d , K_p) can be fixed for desired control performances with respect to a second-order differential equation. The path to follow can simply be defined as the straight line $\Gamma' = (O_{i+1}, \mathbf{Y}_{i+1})$ (refer to ► Fig. 53.5). In this case $c(s) = 0$ and the control law (► 53.2) can be simplified as follows:

$$\delta(y, \theta) = \arctan(-l[\cos^3 \theta(-K_d \tan \theta - K_p y)]) \quad (53.3)$$


The implementation of control law (► 53.3) requires the online estimation of the lateral deviation y and the angular deviation θ of \mathcal{F}_c with respect to Γ . In (Courbon et al. 2009) geometrical relationships between two views are exploited to enable a partial Euclidean reconstruction from which (y, θ) are derived.

5 Example of Results

5.1 Experimental Setup

The experimental vehicle is depicted in  Fig. 53.6. It is an urban electric vehicle, named RobuCab, manufactured by the Robosoft Company. Currently, RobuCab serves as experimental testbed in several French laboratories. The 4 DC motors are powered by lead-acid batteries, providing 2 h autonomy. Vision and guidance algorithms are implemented in C++ language on a laptop using RTAI-Linux OS with a 2 GHz Centrino processor. The Fujinon fisheye lens, mounted onto a Marlin F131B camera, has a field of view of 185°. The image resolution in the experiments was 800×600 pixels. The camera, looking forward, is situated at approximately 80 cm from the ground. The parameters of the rigid transformation between the camera and the robot control frames are roughly estimated. Gray-level images are acquired at a rate of 15 fps. Two illustrative experiments are presented. The first one shows the loop closure performance while the second one shows that it is possible to achieve visual memory-based navigation in large environment.

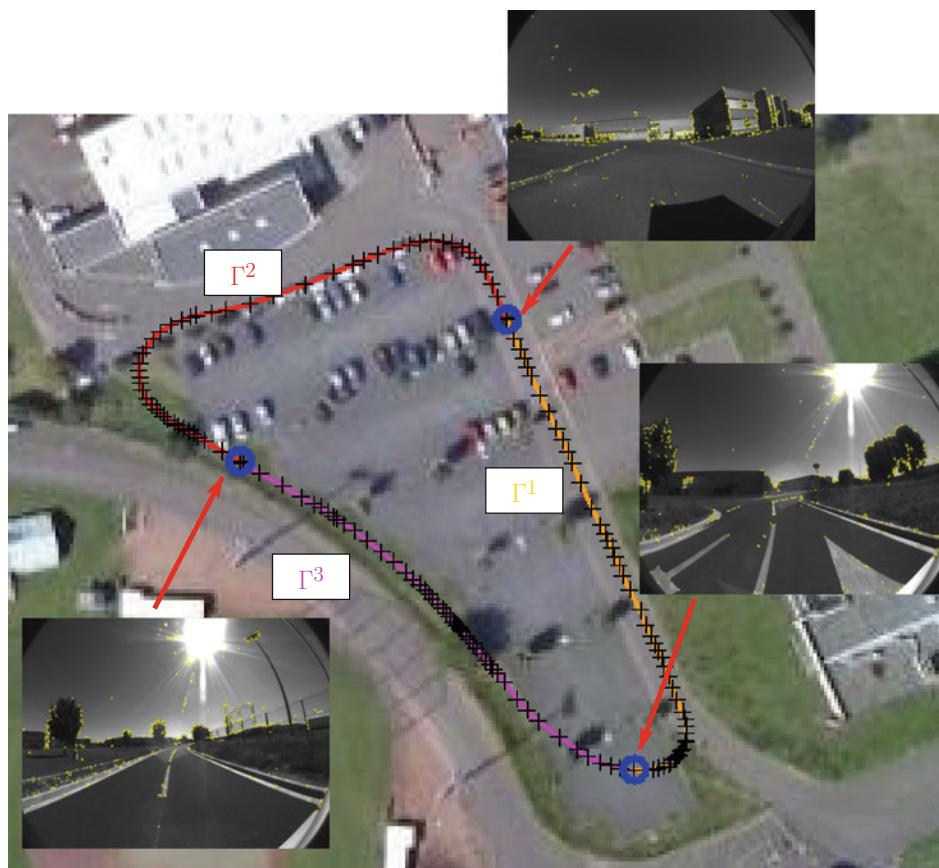


 Fig. 53.6

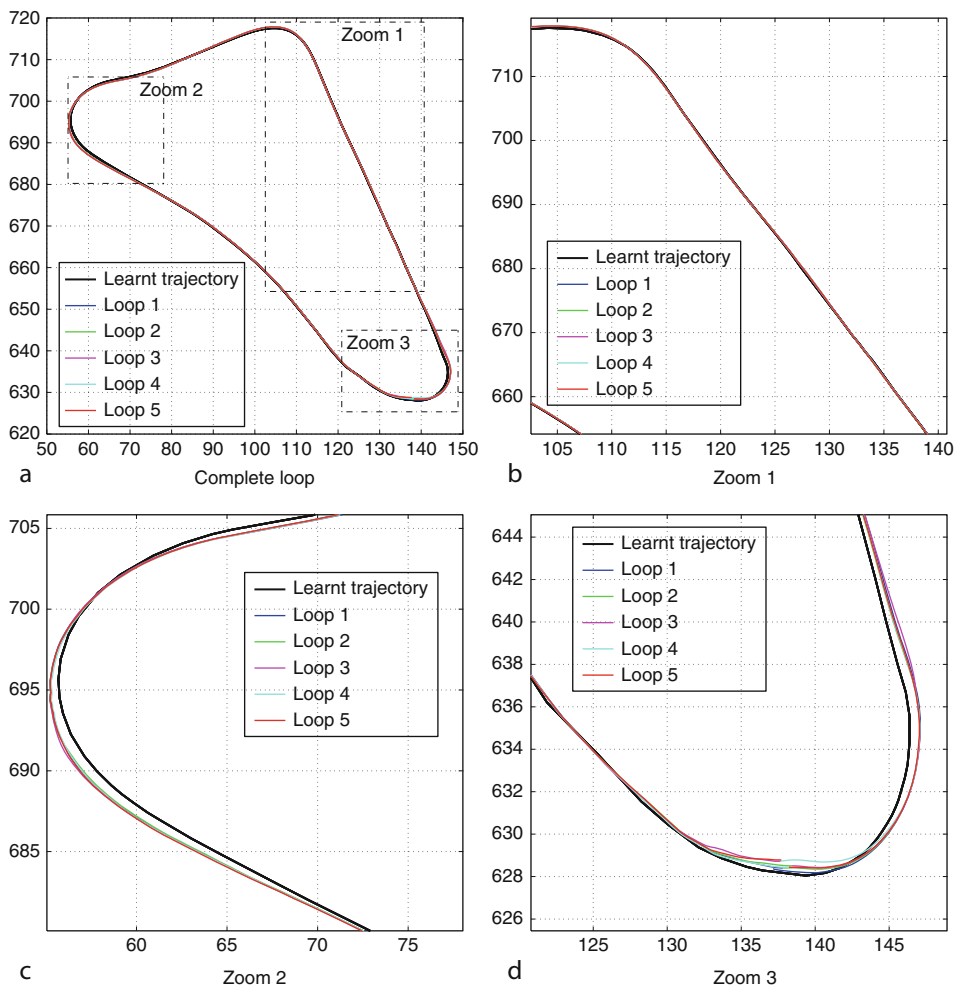
RobuCab vehicle with the embedded camera

5.2 Loop Closure

Autonomous navigation along a loop is interesting because it is a good way to visualize the performances. Remarkably, the topological visual memory implicitly defines the loop closure. It is an advantage of this approach with respect to methods based on a metric representation of the environment which can be subject to significant drift if a loop closure process is not implicitly incorporated to the navigation strategy. The path is defined from the concatenation of the sequences Γ^1 , Γ^2 , and Γ^3 . It is a 270 m loop (refer to ► Fig. 53.7). A total of 1,100 images were acquired and the resulting visual memory contains three sequences and 153 key frames. In this experiment, the navigation task consists in performing five consecutive loops. The results are given in ► Fig. 53.8. One can verify that the robot reaches the position corresponding to the first image of Γ^1 at the end of a “loop.”



► Fig. 53.7
The test Loop



■ Fig. 53.8

The test loop: Trajectory followed during the learning and the autonomous stages

5.3 Large Displacement

This section presents a complete run from path learning to autonomous navigation.

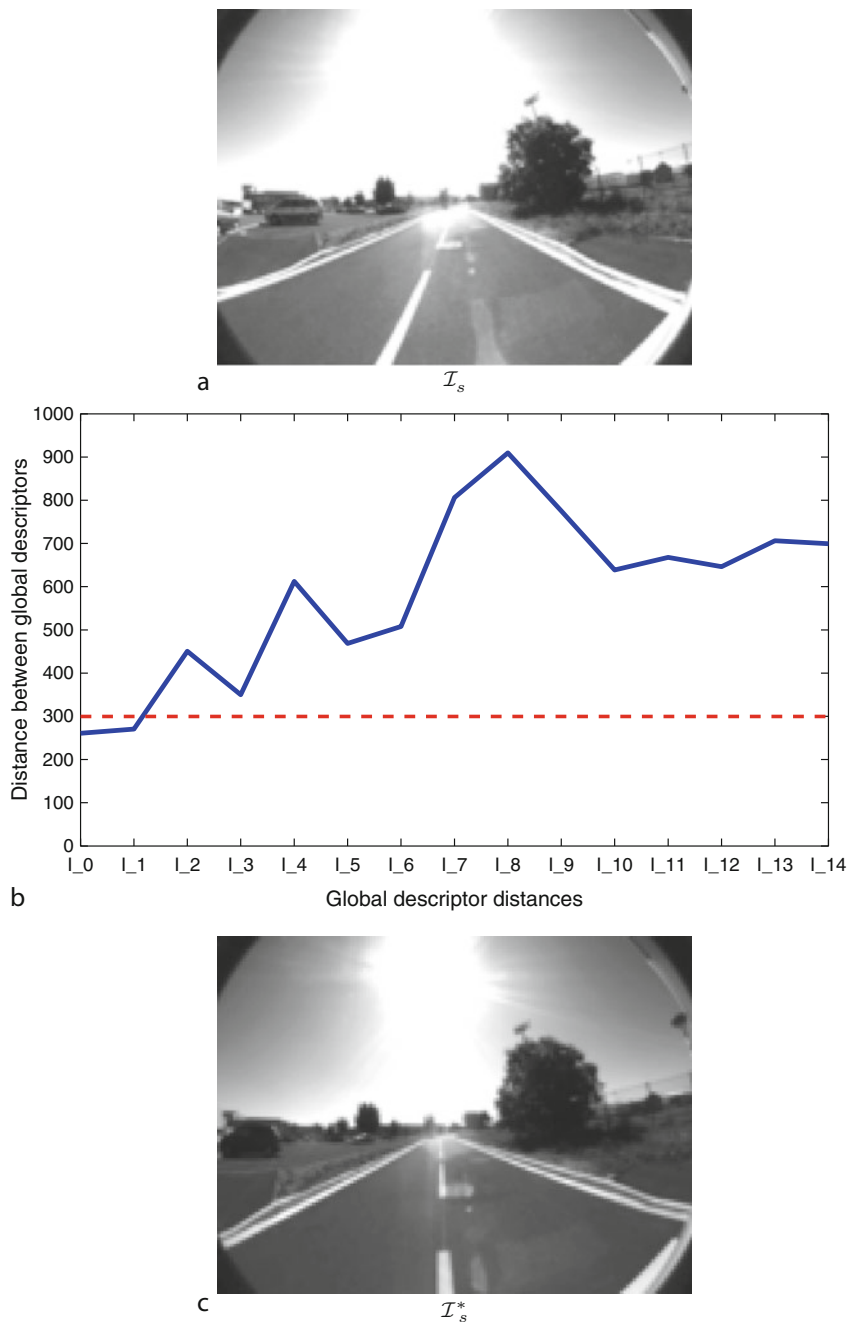
1. *Learning step:* In a second learning phase, the RobuCab was manually driven along the 800-m-long path drawn in blue in ► Fig. 53.11. This path contains important turns as well as way down and up and a comeback. After the selection step, 800 key images are kept and form the visual memory of the vehicle. Some of those images are represented in ► Fig. 53.9.



■ Fig. 53.9

Some key images of the memory

2. *Initial localization*: The navigation task has been started near the visual route to follow (the corresponding image is shown in [Fig. 53.10a](#)). In this configuration, 15 images of the visual memory have been used in the first stage of the localization process. The distances between the global descriptor of the current image and the descriptor of the memorized images (computed offline) are obtained using ZNCC ([Fig. 53.10b](#)). After the second step of the localization process, the image shown in [Fig. 53.10c](#) is chosen as the closest to the image ten (a). Given a goal image, a visual route starting from \mathcal{I}_i^* and composed of 750 key images has been extracted from the visual memory.
3. *Autonomous navigation*: The control ([Fig. 53.3](#)) is used to drive the vehicle along the visual route. A key image is assumed to be reached if the “image error” is smaller than a fixed threshold. In the experiments, the “image error” has been defined as the longest distance (expressed in pixels) between an image point and its position in the desired key image. The longitudinal velocity V is fixed between 1 and 0.4 ms^{-1} . K_p and K_d have been set so that the error presents a double pole located at value 0.3. The vehicle successfully follows the learnt path (refer to [Fig. 53.11](#)). The experiment lasts 13 min for a path of 754 m. A mean of 123 robust matches for each frame has been found. The mean computational time during the online navigation was of 82 ms by image. As can be observed in [Fig. 53.12](#), the errors in the images decrease to zero until reaching a key image. Lateral and angular errors as well as control input are



■ Fig. 53.10

Localization step: I_s is the current initial image. The distance between the current initial image and the key images global descriptors is drawn in (b). After using the local descriptors, I_s^* is selected as the correct image

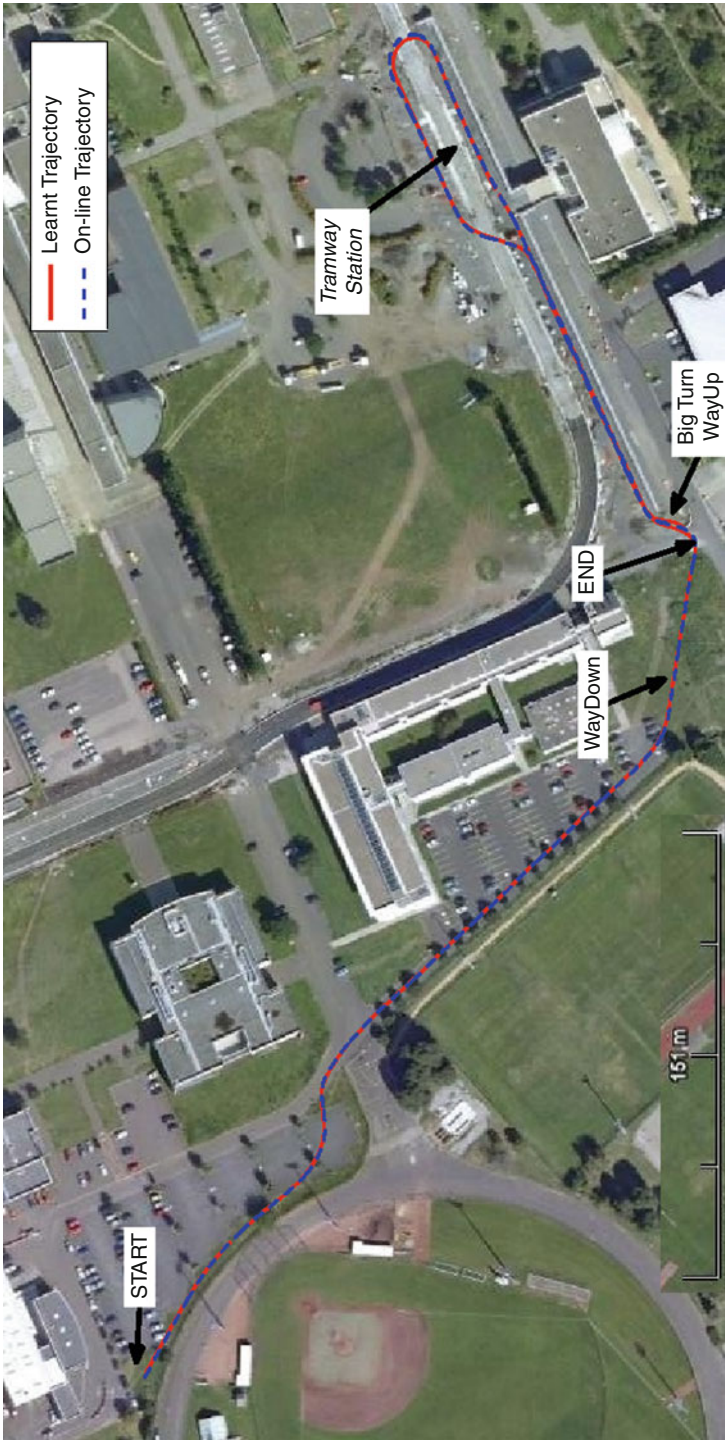
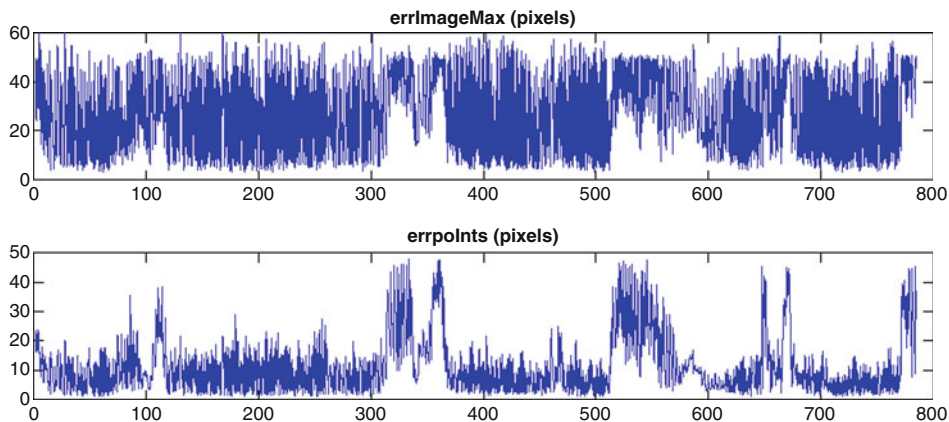
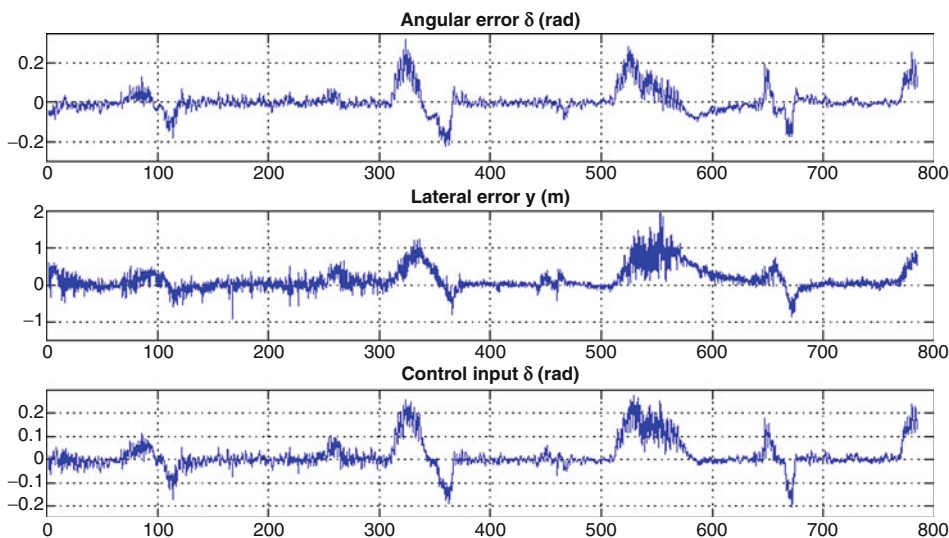


Fig. 53.11 Paths in the university campus executed during the memorization step (in gray) and the autonomous step (in black)



■ Fig. 53.12

Errors in the images (pixels) versus time (second)

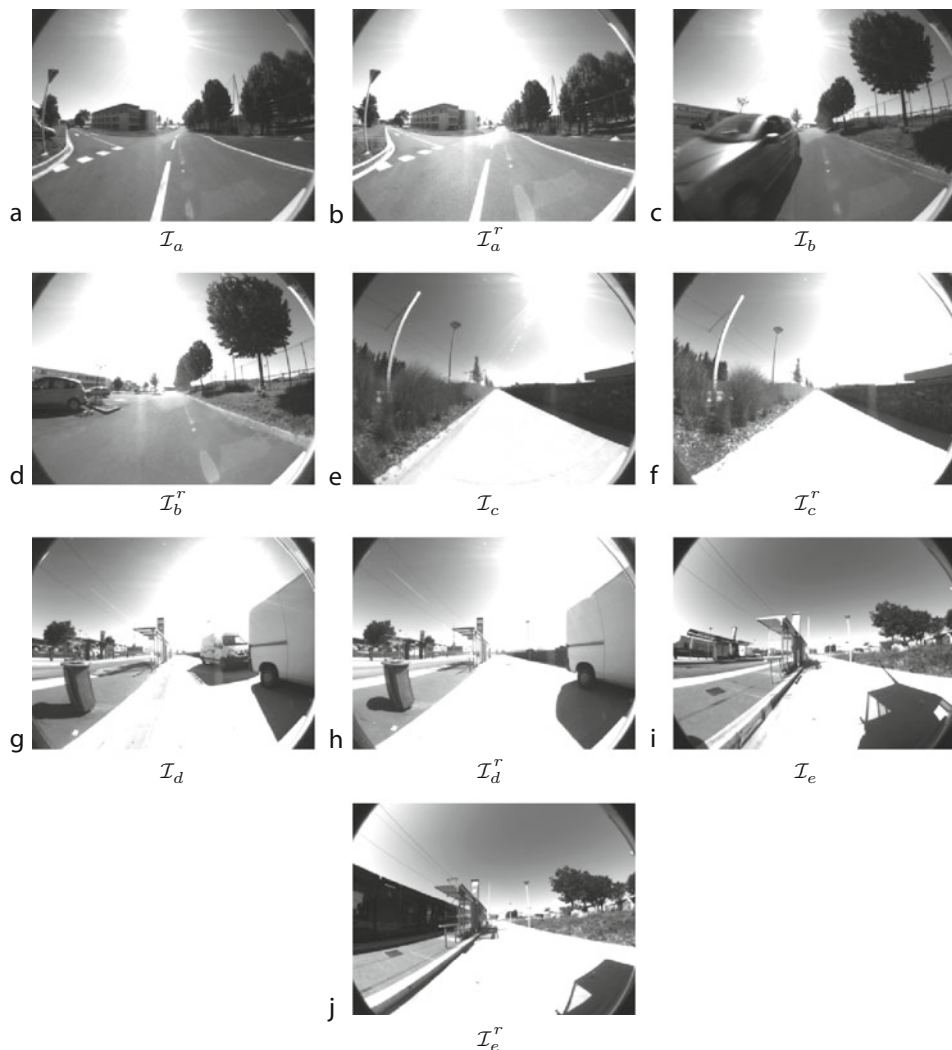


■ Fig. 53.13

Lateral y and angular θ errors and control input δ vs time

represented in [Fig. 53.13](#). As it can be noticed, those errors are well regulated to zero for each key view. Discontinuities due to transitions between two successive key images can also be observed in [Fig. 53.13](#).

Some reached images (with the corresponding images of the memory) are shown in [Fig. 53.14](#). Note that illumination conditions have changed between the memorization and the autonomous steps (refer to [Fig. 53.14a](#) and [b](#) for example) as well as the contents (refer to [Fig. 53.14i](#) and [j](#) where a tram masks many visual features during the autonomous navigation).

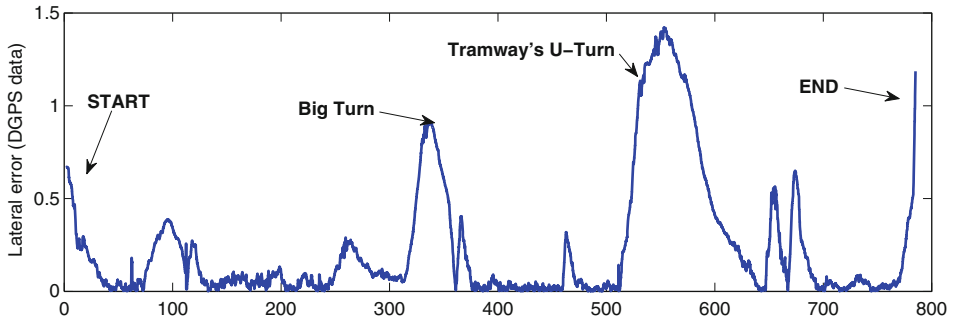


■ Fig. 53.14

Some of the current images \mathcal{I}_k^r where the key images \mathcal{I}_k have been reached

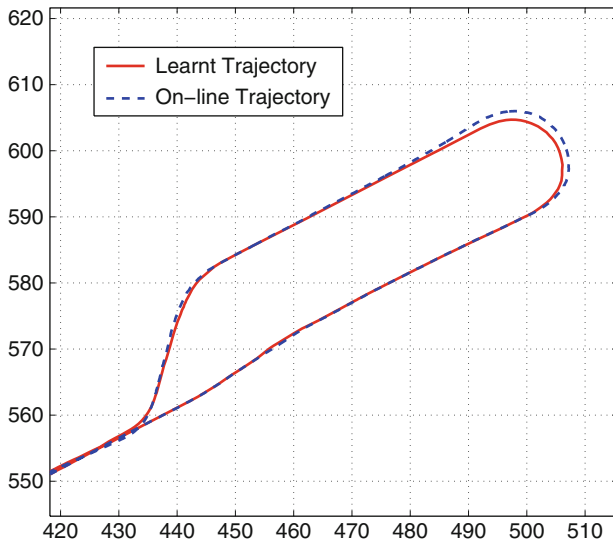
4. *Evaluation with a RTKGPS*: The experimental vehicle has been equipped with a Real-Time Kinematic Differential GPS (Thales Sagitta model). It is accurate to 1 cm (standard deviation) in a horizontal plane when enough satellites are available. The accuracy on a vertical axis is only 20 cm on the hardware platform. The vertical readings are thus discarded and the reported errors are measured in an horizontal plane.

DGPS data have been recorded during the learning and the autonomous stages. The results are reported in ► Fig. 53.15. The red and blue plain lines represent



■ Fig. 53.15

Lateral error (distance between the autonomous and the learnt trajectories, expressed in meter) obtained from DGPS data, vs time



■ Fig. 53.16

Zoom on the trajectories around the tramway station (positions are expressed in m.)

respectively the trajectories recorded during the learning and autonomous stages. It can be observed that these trajectories are similar.

Distances (lateral error) between the vehicle positions during the learning and autonomous stages are reported on [Fig. 53.15](#). The mean of the lateral error is about 25 cm with a standard deviation of 34 cm. The median error is less than 10 cm. The maximal errors are observed along severe turns (see [Fig. 53.16](#) representing a U-turn nearby the tramway station). Note that despite those errors, the visual path is still satisfactory executed (after some images, the vehicle is still at a small distance to the learnt trajectory).

6 Conclusion

This chapter has presented the essential of vision-based topological navigation through an illustrative example. This framework enables a vehicle to follow a visual path obtained during a learning stage using a single camera. The robot environment is modeled as a graph of visual paths, called visual memory from which a visual route connecting the initial and goal images can be extracted. The robotic vehicle can then be driven along the visual route using vision-based control schemes. Importantly, this framework allows loop closure without extra processing.

References

- Andreasson H, Treptow A, Duckett T (2005) Localization for mobile robots using panoramic vision, local features and particle filter. In: IEEE international conference on robotics and automation, ICRA'05, Barcelona, Espagne, Apr 2005, pp 3348–3353
- Andreasson H, Treptow A, Duckett T (2007) Self-localization in non-stationary environments using omni-directional vision. *Robot Auton Syst* 55(7):541–551
- Atkinson R, Shiffrin R (1968) Human memory: a proposed system and its control processes. In: Spence KW, Spence JT (eds) *The psychology of learning and motivation*. Academic, New York
- Bacca B, Salvi J, Battle J, Cufi X (2010) Appearance-based mapping and localization using feature stability histograms. *Electron Lett* 46(16):1120–1121
- Bibby C, Reid I (2007) Simultaneous localisation and mapping in dynamic environments (slamde) with reversible data association. In: *Robotics: science and systems*, Atlanta, GA, USA
- Blaer P, Allen P (2002) Topological mobile robot localization using fast vision techniques. In: IEEE international conference on robotics and automation, ICRA'02, Washington, USA, May 2002, pp 1031–1036
- Chen J, Dixon WE, Dawson DM, McIntire M (2003) Homography-based visual servo tracking control of a wheeled mobile robot. In: *Proceeding of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, Oct 2003, pp 1814–1819
- Courbon J, Mezouar Y, Martinet P (2008) Indoor navigation of a non-holonomic mobile robot using a visual memory. *Auton Robots* 25(3):253–266
- Courbon J, Mezouar Y, Martinet P (2009) Autonomous navigation of vehicles from a visual memory using a generic camera model. *Intell Transport Syst (ITS)* 10:392–402
- Dayoub F, Duckett T (2008). An adaptative appearance-based map for long-term topological localization of mobile robots. In: *IEEE/RSJ international conference on intelligent robots and systems, IROS'08*, Nice, France, Sep 2008, pp 3364–3369
- Dayoub F, Duckett T, Cielniak G (2010) Short- and long-term adaptation of visual place memories for mobile robots. In: *Remembering who we are- human memory for artificial agents symposium*, AISB 2010, Leicester, UK
- DeSouza GN, Kak AC (2002) Vision for mobile robot navigation: a survey. *IEEE Trans Pattern Anal Mach Intell* 24(2):237–267
- Fang Y, Dawson D, Dixon W, de Queiroz M (2002) Homography-based visual servoing of wheeled mobile robots. In: *Conference on decision and control*, Las Vegas, NV, Dec 2002, pp 2866–2871
- Gaspar J, Winters N, Santos-Victor J (2000) Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Trans Robot Autom* 16:890–898
- Goedemé T, Tuytelaars T, Vanacker G, Nuttin M, Gool LV, Gool LV (2005) Feature based omnidirectional sparse visual path following. In: *IEEE/RSJ international conference on intelligent robots and systems*, Edmonton, Canada, Aug 2005, pp 1806–1811

- Gonzalez-Barbosa J, Lacroix S (2002) Rover localization in natural environments by indexing panoramic images. In: IEEE international conference on robotics and automation, ICRA'02, vol 2, Washington, DC, USA, May 2002, pp 1365–1370
- Harris C, Stephens M (1988) A combined corner and edge detector. In: The fourth alvey vision conference, Manchester, UK, pp 147–151
- Hayet J, Lerasle F, Devy M (2002) A visual landmark framework for indoor mobile robot navigation. In: IEEE international conference on robotics and automation, ICRA'02, Washington, DC, USA, May 2002, pp 3942–3947
- Hochdorfer S, Schlegel C (2009) Towards a robust visual SLAM approach: addressing the challenge of life-long operation. In: 14th international conference on advanced robotics, Munich, Germany
- Ieng S, Benosman R, Devars J (2003) An efficient dynamic multi-angular feature points matcher for catadioptric views. In: Workshop OmniVis'03, in conjunction with computer vision and pattern recognition (CVPR), vol 07, Wisconsin, USA, Jun 2003, p 75
- Jones S, Andresen C, Crowley J (1997) Appearance-based process for visual navigation. In: IEEE/RSJ international conference on intelligent robots and systems, IROS'97, vol 2, Grenoble, France, pp 551–557
- Lemaire T, Berger C, Jung I, Lacroix S (2007) Vision-based slam: stereo and monocular approaches. *Int J Comput Vision* 74(3):343–364
- Linåker F, Ishikawa M (2004) Rotation invariant features from omnidirectional camera images using a polar higher-order local autocorrelation feature extractor. In: IEEE/RSJ international conference on intelligent robots and systems, IROS'04, vol 4, Sendai, Japon, Sep 2004, pp 4026–4031
- Lowe D (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
- Ma Y, Kosecka J, Sastry SS (1999) Vision guided navigation for a nonholonomic mobile robot. *IEEE Trans Robot Autom* 15(3):521–537
- Matsumoto Y, Inaba M, Inoue H (1996) Visual navigation using view-sequenced route representation. In: IEEE international conference on robotics and automation, ICRA'96, vol 1, Minneapolis, Minnesota, USA, Apr 1996, pp 83–88
- Matsumoto Y, Ikeda K, Inaba M, Inoue H (1999) Visual navigation using omnidirectional view sequence. In: IEEE/RSJ international conference on intelligent robots and systems, IROS'99, vol 1, Kyongju, Corée, Oct 1999, pp 317–322
- Menegatti E, Zoccarato M, Pagello E, Ishiguro H (2003) Hierarchical image-based localisation for mobile robots with monte-carlo localisation. In: European conference on mobile robots, ECMR'03, Varsovie, Pologne, Sep 2003
- Mouragnon E, Lhuillier M, Dhome M, Dekeyser F, Sayd P (2009) Generic and real-time structure from motion using local bundle adjustment. *Image Vision Comput* 27(8):1178–1193
- Murillo A, Guerrero J, Sagiés C (2007) SURF features for efficient robot localization with omnidirectional images. In: IEEE international conference on robotics and automation, ICRA'07, Rome, Italie, Apr 2007, pp 3901–3907
- Nistér D (2004) An efficient solution to the five-point relative pose problem. *Trans Pattern Anal Mach Intell* 26(6):756–770
- Pajdla T, Hlaváč V (1999) Zero phase representation of panoramic images for image based localization. In: 8th international conference on computer analysis of images and patterns, Ljubljana, Slovénie, Sep 1999, pp 550–557
- Pollefeys M, Gool LV, Vergauwen M, Verbiest F, Cornelis K, Tops J, Koch R (2004) Visual modeling with a hand-held camera. *Int J Comput Vision* 59(3):207–232
- Royer E, Lhuillier M, Dhome M, Lavest J-M (2007) Monocular vision for mobile robot localization and autonomous navigation. *Int J Comput Vision* 74:237–260, special joint issue on vision and robotics
- Samson C (1995) Control of chained systems. Application to path following and time-varying stabilization of mobile robots. *IEEE Trans Autom Control* 40(1):64–77
- Svoboda T, Pajdla T (2001) Matching in catadioptric images with appropriate windows and outliers removal. In: 9th international conference on computer analysis of images and patterns, Berlin, Allemagne, Sep 2001, pp 733–740
- Tamimi A, Andreasson H, Treptow A, Duckett T, Zell A (2005) Localization of mobile robots with omnidirectional vision using particle filter and iterative SIFT. In: 2nd European conference

- on mobile robots (ECMR), Ancona, Italie, Sep 2005, pp 2–7
- Thormählen T, Broszio H, Weissenfeld A (2004) Keyframe selection for camera motion and structure estimation from multiple views. In: 8th European conference on computer vision (ECCV), Prague, Czech Republic, May 2004, pp 523–535
- Thuilot B, Bom J, Marmoiton F, Martinet P (2004) Accurate automatic guidance of an urban electric vehicle relying on a kinematic GPS sensor. In: 5th IFAC symposium on intelligent autonomous vehicles, IAV'04, Instituto Superior Técnico, Lisbonne, Portugal, Jul 2004
- Torr P (2002) Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *Int J Comput Vision* 50(1):35–61
- Triggs B, McLauchlan P, Hartley R, Fitzgibbon A (2000) Bundle adjustment – a modern synthesis. In: Triggs B, Zisserman A, Szeliski R (eds) *Vision algorithms: theory and practice*, vol 1883, Lecture notes in computer science. Springer, Berlin, pp 298–372
- Tsakiris D, Rives P, Samson C (1998) Extending visual servoing techniques to nonholonomic mobile robots. In: GHD Kriegman, A Morse (eds) *The confluence of vision and control*, LNCIS, vol 237. Springer, London/New York, pp 106–117
- Wangsiripitak S, Murray D (2009) Avoiding moving outliers in visual SLAM by tracking moving objects. In: IEEE international conference on robotics and automation, ICRA'09, Kobe, Japan, pp 705–710
- Yamauchi B, Langley P (1997) Spatial learning for navigation in dynamic environments. *IEEE Trans Syst Man Cybern* 26(3):496–505
- Zhang Z, Deriche R, Faugeras O, Luong Q-T (1995) A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artif Intell J* 78:87–119
- Zodiac T (1995) In: dewit Canedas C, Siciliano B, Bastin G (eds) *Theory of robot control*. Springer, Berlin

54 Awareness of Road Scene Participants for Autonomous Driving

*Anna Petrovskaya*¹ · *Mathias Perrollaz*² · *Luciano Oliveira*³ · *Luciano Spinello*^{4,8} · *Rudolph Triebel*^{5,8} · *Alexandros Makris*⁹ · *John-David Yoder*⁶ · *Christian Laugier*² · *Urbano Nunes*³ · *Pierre Bessiere*⁷

¹Artificial Intelligence Laboratory, Stanford University, Stanford, CA, USA

²e-Motion Project-Team, INRIA Grenoble Rhône-Alpes, Saint Ismier Cedex, France

³Faculty of Science and Technology, Coimbra University, Pólo II, Coimbra, Portugal

⁴University of Freiburg, Freiburg, Germany

⁵University of Oxford, Oxford, UK

⁶Mechanical Engineering Department, Ohio Northern University, Ada, OH, USA

⁷INRIA and Collège de France, Saint Ismier Cedex, France

⁸Inst.f. Robotik u. Intelligente Systeme, ETH Zurich, Zurich, Switzerland

⁹INRIA, Saint Ismier Cedex, France

1	Introduction	1385
1.1	Taxonomy of DATMO Applications	1386
1.2	DATMO Problem Statement	1387
1.2.1	Coordinate Frames	1387
1.2.2	Bayesian Formulation of DATMO	1387
2	Sensors and Models	1389
2.1	Sensors	1389
2.1.1	Optical Cameras	1389
2.1.2	Laser Range Finders	1390
2.1.3	Radars	1391
2.2	Models of Sensors and Motion	1392
2.2.1	Physical Models for Range Sensors	1393
2.2.2	Pseudo-Position Sensors	1393

- 2.2.3 Virtual Sensors 1393
- 2.2.4 Dynamics Models 1396
- 2.3 Scene Models 1396
 - 2.3.1 Cluster-Based Models 1396
 - 2.3.2 Geometric Models 1398
 - 2.3.3 Grid-Based Models 1399
- 3 *Inference: Filtering and Tracking* 1400
 - 3.1 Track Management and Dynamics Update 1402
 - 3.1.1 Imposing Track Constraints 1402
 - 3.1.2 Track Existence 1402
 - 3.1.3 Track Creation 1403
 - 3.1.4 Track Deletion 1403
 - 3.1.5 Dynamics Update 1403
 - 3.2 Traditional DATMO 1404
 - 3.2.1 Data Segmentation 1404
 - 3.2.2 Data Association 1404
 - 3.3 Model-Based DATMO 1406
 - 3.3.1 Rao-Blackwellization 1406
 - 3.3.2 Scaling Series Algorithm 1408
 - 3.3.3 MCMC Approach 1408
 - 3.4 Grid-Based DATMO 1409
 - 3.4.1 The Bayesian Occupancy Filter 1410
 - 3.4.2 BOF Implementation 1412
 - 3.4.3 Object Level Tracking 1413
- 4 *Pattern Recognition and Sensor Fusion* 1414
 - 4.1 Pattern Recognition 1414
 - 4.1.1 Vision-Based approaches 1414
 - 4.1.2 Object and People Detection from 2D Laser Data 1418
 - 4.1.3 Example of Detection by Classifier Ensemble 1420
 - 4.2 Detection by Sensor Fusion 1422
- 5 *Conclusions* 1428

Abstract: This chapter describes detection and tracking of moving objects (DATMO) for purposes of autonomous driving. DATMO provides awareness of road scene participants, which is important in order to make safe driving decisions and abide by the rules of the road. Three main classes of DATMO approaches are identified and discussed. First is the traditional approach, which includes data segmentation, data association, and filtering using primarily Kalman filters. Recent work within this class of approaches has focused on pattern recognition techniques. The second class is the model-based approach, which performs inference directly on the sensor data without segmentation and association steps. This approach utilizes geometric object models and relies on non-parametric filters for inference. Finally, the third class is the grid-based approach, which starts by constructing a low level grid representation of the dynamic environment. The resulting representation is immediately useful for determining free navigable space within the dynamic environment. Grid construction can be followed by segmentation, association, and filtering steps to provide object level representation of the scene. The chapter introduces main concepts, reviews relevant sensor technologies, and provides extensive references to recent work in the field. The chapter also provides a taxonomy of DATMO applications based on road scene environment and outlines requirements for each application.

1 Introduction

Autonomous driving in populated areas requires great situational awareness. The autonomous vehicle must perceive not only the stationary environment, but also dynamic objects such as vehicles and pedestrians. For each moving target, the autonomous vehicle needs to identify location and velocity, so that it can predict the target's position a few seconds later for planning purposes. Awareness of moving objects includes both detection of new targets and tracking of existing targets over time. For this reason, it is often referred to as *detection and tracking of moving objects* or DATMO for short.

The need for DATMO was born with the first interest in intelligent and autonomous vehicles in 1980s. A large exploratory project was launched in Europe in 1986 under the name PROMETHEUS, followed by a number of initiatives in Japan and United States (Bertozzi et al. 2000; Sun et al. 2006). Fueled by innovations in sensor technologies, recent DATMO advances have focused on improved detection using pattern recognition techniques, increased robustness using sensor fusion, and more accurate modeling of sensors and the scene. This chapter gives an overview of the most prominent DATMO approaches and relevant concepts.

The remainder of this section provides a taxonomy of DATMO applications and gives a formal description of the DATMO problem. ➤ [Section 2](#) introduces the required concepts as well as describes sensors and models. ➤ [Section 3](#) discusses inference techniques for DATMO. In particular, it outlines three main classes of DATMO approaches: traditional, model based, and grid based. ➤ [Section 4](#) covers pattern recognition, and sensor fusion. The chapter concludes with a discussion in ➤ [Sect. 5](#).

1.1 Taxonomy of DATMO Applications

Although detection of people and vehicles has received a lot of attention in other fields, this chapter focuses on techniques that satisfy the high demands of autonomous driving. To illustrate these demands, the driving applications are roughly grouped into three categories: (1) pedestrian zone driving, (2) freeway driving, and (3) mixed urban driving.

In the first category, the robot operates in a pedestrian zone, which can be crowded by bicyclists, adults, children, and pets. The robot has to operate in close proximity to these inhabitants in a safe and courteous manner. This situation is challenging because the targets come in all shapes and sizes, can change direction of motion unpredictably, and can partially occlude themselves or each other. The operational velocity has to be similar to pedestrian speed: 1–2 m/s. Since the operational velocities are relatively low, the required range of target detection and tracking is relatively short: a distance of 10–20 m is usually sufficient for safe operation. On the other hand, due to close proximity to targets, the robot's reaction time has to be similar to human reaction time, which is usually taken to be 1 s. Hence, detection and tracking have to happen fast enough, so that the entire perception-planning-control loop can be executed at 1 Hz. However, to simplify the association stage, tracking at 5–10 Hz is desirable.

In the second category, the robot drives on a freeway, which is a much more structured environment than a pedestrian zone. The environment is populated only by vehicles, which have to travel within lanes most of the time. Oncoming vehicles are confined to a separate part of the freeway, so the robot only needs to concern itself with vehicles moving in the same direction. The motion of these vehicles is governed by non-holonomic constraints of vehicle dynamics and therefore is much more predictable than motion of pedestrians. The main challenge in this environment comes from high operational velocities: up to 35 m/s. High operational velocity leads to longer stopping distances. Thus, the required range for detection and tracking of targets is much greater than in the first category: 100 m or more. In order to plan smooth maneuvers at high speeds, the planning loop runs at 10–20 Hz and hence tracking also needs to happen at a comparable rate.

The third category lies in between the first two. The robot operates in an urban setting with mixed pedestrian and vehicle traffic. This category combines challenges of the first two categories. The environment is not as structured as on the freeway – vehicles and pedestrians can cross the robot's path in unpredictable ways. Oncoming traffic is usually present, which doubles the effective operational velocity from the perspective of object detection. If the speed limit is 15 m/s (35 mph), then an oncoming vehicle can move with a velocity of up to 30 m/s with respect to the robot. Hence, the detection range and tracking frequency rate have to be almost as high as on freeways: 60–80 m and 10 Hz respectively. In addition, the robot has to be able to distinguish between different types of targets: pedestrians, bicyclists, and vehicles. These three types of targets are governed by different traffic and dynamic laws, and hence, the robot needs to treat them differently.

1.2 DATMO Problem Statement

In DATMO, an autonomous vehicle (also called *robot* or *ego-vehicle*) navigates in a populated outdoor environment. Sensors are mounted on the ego-vehicle itself, which is potentially moving at high speeds. The robot is to perceive the environment around it based on its sensors and to detect and track all moving objects in its vicinity. For high level decision making, the robot needs to estimate pose and velocity of each moving object based on the sensor measurements. For low level trajectory planning, the robot needs to estimate the free space for driving.

Object pose estimates (also called *tracks*) need to be produced at a high enough rate to be useful for the perception-planning-control loop, which typically runs at 1–10 Hz. *False negatives* (i.e., missing an object) tend to be very dangerous, whereas *false positives* (i.e., finding an object when one is not there) are slightly more acceptable. Note that for the applications in *advanced driver assistance systems* (ADAS), it is the opposite.

1.2.1 Coordinate Frames

This chapter assumes that a reasonably precise pose of the robot is always available. Further, it assumes the use of *smooth coordinates*, which provide a locally consistent estimate of the ego-vehicle motion. Smooth coordinates should not experience sudden jumps because jumps can greatly increase tracking uncertainty. In practice, smooth coordinates can be obtained by integrating the ego-vehicle velocity estimates from the inertial navigation system (Montemerlo et al. 2008; Leonard et al. 2008). To map from smooth coordinates to globally consistent GPS coordinates, one simply needs to add an offset, which is periodically updated to reflect the mismatch between the smooth and GPS coordinate systems. In the remainder of this chapter, all operations are carried out in the smooth coordinate frame, which will also be called the *world frame*. The transformation from smooth to GPS coordinates is only needed when dealing with global features, such as the digital road map.

It is also common to use local coordinate systems, which are tied to the robot, the sensor, or tracked objects. These are called the *robot coordinate frame*, the *sensor coordinate frame*, and the *object coordinate frame* respectively.

1.2.2 Bayesian Formulation of DATMO

For a general Bayesian problem, the goal is to infer the state \mathbf{S} of a system (or of the world) based on a set of measurements \mathbf{Z} . Due to uncertainty, this information is best captured as a probability distribution $bel := p(\mathbf{S}|\mathbf{Z})$ called the *posterior distribution* or the *Bayesian belief*.

In a dynamic Bayesian system, the state changes over time, which is assumed to be discretized into small (although not necessarily equal) time intervals. The system is assumed to evolve as a *Markov process* with unobserved states. The goal is to estimate the belief at time t , $bel_t := p(\mathbf{S}_t|\mathbf{Z}_1, \dots, \mathbf{Z}_t)$. The behavior of the system is described via two

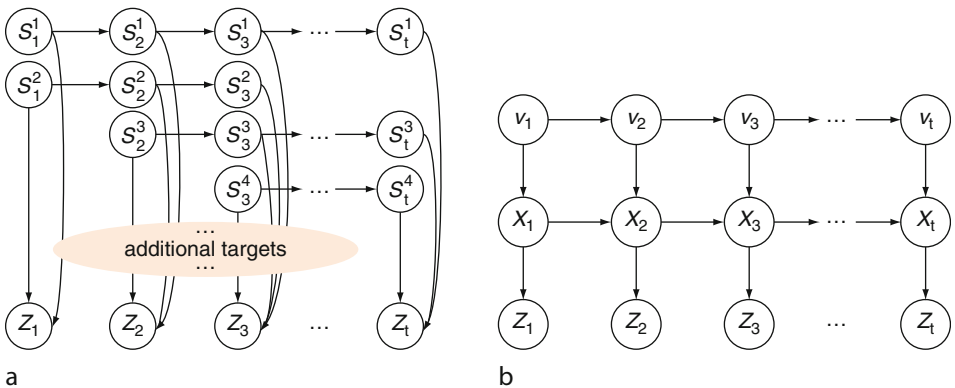
probabilistic laws: (i) the *measurement model* $p(Z_t|S_t)$ captures how the sensor measurements are obtained and (ii) the *dynamics model* $p(S_t|S_{t-1})$ captures how the system evolves between time steps. Given these two laws and measurements up to time t , the belief can be computed recursively using a *Bayesian filter* algorithm, which relies on the *Bayesian recursion equation*:

$$bel_t = \eta p(Z_t|S_t) \int p(S_t|S_{t-1}) bel_{t-1} dS_{t-1}, \quad (54.1)$$

where η denotes the normalization constant.

For the DATMO problem, the system consists of a set of moving targets T^1, \dots, T^{K_t} . The number of targets K_t changes over time as some targets leave and others enter the scene. For each target, the estimated parameters include its momentary 2D pose $X_t = (x_t, y_t, \theta_t)$ consisting of its 2D position (x_t, y_t) and orientation θ_t at time t . The parameters also include each target's forward velocity v_t , which is typically a scalar as the objects are assumed to move along vectors aligned with their orientation. Thus, the Bayesian state for a single target is usually $S_t := (X_t, v_t)$, although in some cases, additional parameters may be added to describe shape, class, or motion of objects. The full Bayesian state consists of the states of all the targets: $S_t := (S_t^1, \dots, S_t^{K_t})$. At each time step, we obtain a new sensor measurement Z_t .

A graphical representation of the resulting probabilistic model for DATMO is shown in [Fig. 54.1](#). Note that during filtering, the targets are usually considered independent of each other, although some relationships clearly exist in reality. The independence assumption allows the problem to be split into multiple smaller sub-problems: one per target. This reduces the effective dimensionality that the estimation algorithms need to cope



■ Fig. 54.1

Graphical representation of the probabilistic model for DATMO. (a) For the full DATMO problem, the graphical model shows a variable number of targets. (b) For a single target, the graphical model shows the relationships between the target's pose X_t , forward velocity v_t , and measurements Z_t

with. Relationships between targets are often introduced as a set of constraints, which are imposed after each filtering step as we discuss in [Sect. 3.1](#).

For a single target, the dependencies between the parameters are typically modeled via several probabilistic laws: the *velocity model* $p(v_t|v_{t-1})$, the *motion model* $p(X_t|X_{t-1}, v_t)$, and the measurement model $p(Z_t|X_t)$. The velocity and motion models together comprise the dynamics model. Measurement and dynamics models for DATMO are discussed in [Sect. 2.2](#).

2 Sensors and Models

2.1 Sensors

The most common sensors used in DATMO approaches are *optical cameras* and *laser range finders* although some systems also incorporate *radar* sensors. In this subsection, we discuss all three types of sensors, their operating characteristics, advantages, and limitations.

2.1.1 Optical Cameras

Cameras are the most popular sensors due to their low cost and high information content. Two major camera technologies are available: *charge-coupled devices* (CCD) and *complementary metal oxide semiconductors* (CMOS). CCD cameras tend to have a higher output uniformity because all pixels can be devoted to capture light. In contrast, CMOS cameras are usually more compact and require less off-chip circuitry. An important parameter for cameras is their *field of view* (FOV), which is directly defined by the optics used and by the size of the sensor's matrix. A very narrow field of view is achieved with a *tele-lens* and a very wide field of view results from using a *fish-eye lens*. While in the first case, far objects can be observed at a higher resolution and nearly no line distortion, in the second case, a much larger fraction of the environment can be observed within a single image.

Cameras are very attractive sensors because they capture high-resolution images of the environment at high frame rates, while consuming very little energy. These sensors produce high volumes of data with high information content. While this is an important advantage over other sensors, high volumes of data also lead to significant challenges in transmission and processing of the data stream. Moreover, since cameras do not capture range to objects, the data produced by cameras are more difficult to interpret than range finder data. Cameras are also greatly impacted by changes in lighting conditions, shadows, and other types of ambient effects. Finally, unlike radars and lasers, cameras do not provide useful data after dark, when the lighting of the scene is insufficient.

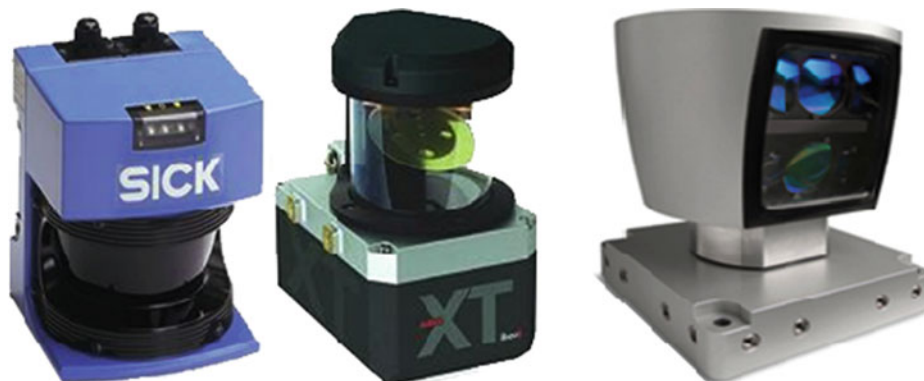
In addition to ordinary *monocular cameras* several special camera types and configurations are employed. *Stereo cameras* make use of two (or more) cameras for binocular vision, which is able to produce range data in addition to ordinary monocular images. This is a cost-effective sensor for acquiring range data, but it can be very sensitive to calibration errors. One popular established stereo system is the Bumblebee camera by

Point Grey. Omni-directional cameras capture 360° view of the environment. They are constructed either by using multiple monocular cameras (e.g., the Ladybug camera) or mirrors. The multiple camera approach leads to even greater volumes of data, whereas the mirror approaches can lead to a loss of resolution. Infrared cameras perceive thermal energy emitted by objects in the environment. Although these sensors can simplify detection of people, the signal-to-noise ratio is very high and the range is limited to short distances.

2.1.2 Laser Range Finders

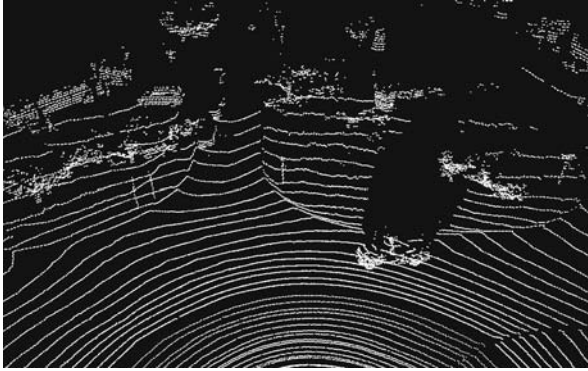
Laser range finders retrieve distances by measuring *time-of-flight* of laser beam returns. Several types are available on the market: 2D, multilayer 2D, and 3D (► Fig. 54.2). The principle of operation of 2D lasers is based on a rotating mirror that deflects a beam into a plane. The reflected beam is detected by a receiver circuit that also computes the traveled distance using the time elapsed from beam emission to detection. A very wide field of view is retrieved in this manner (at least 120°), but the scan only captures information about a 2D slice of the world. Measurements are very precise, but sparse due to limited angular resolution (0.25–1°). Laser range finders work in any lighting conditions, although direct sunlight can cause false measurements and errors. For DATMO applications, the 2D lasers are typically mounted horizontally on the ego-vehicle. This produces scan lines parallel to the ground. A significant challenge is to tell the difference between ground readings vs. readings obtained from true obstacles.

Recently lasers have evolved from 2D to multilayer 2D devices, which are more suitable for driving applications. The principle of operation is similar to a standard 2D scanner, but more slices are retrieved (slices are almost parallel 1–2°) by employing more beams and more receivers. For example, IBEO Alasca sensors allow for easy ground filtering by collecting four nearly parallel horizontal scan lines. Using information from all four scan lines, it is easier to



■ Fig. 54.2

Laser range finders. From left to right: 2D laser (Sick LMS200), multilayer 2D laser (IBEO Alasca XT), 3D laser (Velodyne HDL-64E)



■ Fig. 54.3

Example of a 3D scan obtained with a Velodyne HDL-64E scanner. See ► Fig. 54.4 for a picture of the scene

tell which readings likely came from the ground, by considering the vertical alignment of impacts and, hence, these sensors can provide automatic ground filtering.

Another recent innovation are 3D laser range finders, e.g., the Velodyne HDL-64E sensor. Data are retrieved at 5–15 Hz with up to two million points per second (see ► Fig. 54.3). Even though angular resolution is not as fine as for a camera, this sensor has all the aforementioned advantages of a range device. This sensor has 64 lasers aligned at different vertical angles on a cylinder rotating around a vertical axis. On each rotation, lasers sweep the space and a 3D point cloud is received. The first successful usage of this device in the field of vehicle detection was presented at the DARPA Urban Challenge in 2007. Given the rich data produced by 3D scanners, the challenge has become to process the readings in real time, as target tracking at 10–20 Hz is desirable for driving decision making.

Overall, lasers can provide data at high frame rates and with high information content (especially 3D lasers). Unlike cameras, lasers give range information for every data point and can work in the dark. However, compared to cameras, lasers are much more expensive and consume more energy, which is required for sending out laser light pulses. Their angular resolution is not as high as cameras and they do not provide color information. Compared to radar, lasers can be affected by weather conditions, such as rain or fog.

2.1.3 Radars

Radar sensors measure distance to obstacles by sending out radio waves and measuring either time delay or phase shift of the reflected signals. The typical radars used for autonomous driving applications are *millimeter wave (MMW)* radars, which have started to appear on some upscale commercial vehicles for *adaptive cruise control*.

Although radar speed guns are commonly used by law enforcement officers to detect velocity of speeding vehicles, this technology is more difficult to use on an autonomous

platform. Radars tend to have narrow field of view and low resolution for position measurements, although they are very precise for velocity estimation. The measured range can be corrupted by echoes from multiple objects in the path of the wave and even by highly reflective objects outside of the FOV. Some targets get multiple detections, others go undetected, and the bearing to each target is not very accurately determined. For these reasons, radars are often combined with vision sensors to obtain better accuracy (Lundquist and Schon 2008; Richter et al. 2008). Since static clutter and infrastructure cause echoes, it is common to filter out static returns from radars, making these sensors better suited for highway rather than urban applications.

Despite all the challenges, radars perform better than other sensors in adverse weather conditions and have longer detection range. Hence, these sensors warrant further evaluation for the purposes of DATMO.

2.2 Models of Sensors and Motion

Measurement models (also called *sensor models* or *observation models*) define the measurement process in probabilistic terms. In other words, given a pose X_t of a tracked object, the measurement model defines the probability of obtaining a specific set of sensor measurements Z_t . Of course, it is not possible to model the sensor operation exactly as many unknown effects can impact the performance of the sensor: lighting conditions, dust, or fog, for example. Hence, measurement models explicitly represent some properties of the sensors and consider the un-modeled effects as uncertainty or noise.

Measurement models in DATMO fall within a spectrum between *physical sensor models* and *pseudo-position models*. The first type attempts to model the physical sensor operation directly by considering the physical phenomena that occur when the sensor interacts with the world (e.g., laser rays travel through space and get reflected from objects). These methods are appropriate when the physical phenomena are easy to model (i.e., usually for range sensors rather than classical vision). Examples of physical sensor models are discussed in 🔗 Sect. 2.2.1 below.

On the opposite side of the spectrum are methods, which derive target positions (and sometimes orientations and velocities) from sensor data prior to the application of the Bayesian filter. Then the probabilistic sensor model within the Bayesian filter represents the resulting *pseudo-sensor* of target positions (🔗 Sect. 2.2.2).

Most sensor model approaches will use at least some data preprocessing techniques to enhance the data. When the resulting data are still far from the quantities to be estimated, a lot of work still remains for DATMO algorithms. The sensor data at this stage shall be called *virtual sensor* data. Virtual sensor techniques can be very light (e.g., apply a Gaussian blur) or they can build a qualitatively different sensor (e.g., stereovision). The important distinction from pseudo-position sensors is that virtual sensors are still *relative sensors* in that they do not provide direct measurements of quantities to be estimated. Virtual sensor techniques are discussed in 🔗 Sect. 2.2.3.

2.2.1 Physical Models for Range Sensors

Physical sensor models are most common for range sensors. Two main types are prevalent: the *proximity model* and the *independent beam model*. In the proximity model, each range measurement is converted into a 3D point in the world frame. These points are considered to be caused by objects in the environment (both static and dynamic). The closer the points are to the surface of the objects the more likely the measurements are, hence the name proximity model. Proximity models are fast to compute, but they discard *negative information*, i.e., the information that the space the laser rays traveled through must be unoccupied. This information is taken into account by the *independent beam model (IB)*. In the IB model, ray tracing is used to identify range to the closest obstacle along each laser ray. The obtained range is compared to the range returned by the sensor. The closer together these range values the more likely the measurements. Both of these models have been used with great success in SLAM, where proximity models are often called *likelihood fields* (Thrun et al. 2005).

2.2.2 Pseudo-Position Sensors

Most of the sensors used for DATMO provide only relative information about the targets, e.g., individual range measurements or pixels. The quantities that need to be estimated are positions, orientations, and velocities of targets. Many DATMO approaches augment the physical sensors with a set of algorithms (as discussed in ► Sect. 3.2) to produce pseudo-measurements of target positions (and in some cases orientations and velocities). The resulting pseudo-measurements are modeled as direct measurements of target positions, corrupted by zero-mean Gaussian noise.

2.2.3 Virtual Sensors

It is often useful to augment the physical sensor by a number of data preprocessing or clean-up techniques. A variety of methods can be used.

For range data, it is common to sub-sample the rays, readjust origin point of rays, project from 3D to 2D, and reject outliers. It is especially important to filter out ground readings. ► Figure 54.4 shows an example of a virtual sensor, where a 3D scan is converted to a 2D scan. With multilayer and 3D laser sensors, ground filtering can be done based on the fact that impacts are aligned almost vertically on objects. Hence, ground filtering can be achieved by comparing relative angles between rays (see ► Fig. 54.5 for an example). When a vision sensor is available in addition to laser, the vision sensor can be used to detect the road surface, thus allowing to filter range data.

Since lasers do not see black obstacles very well, some approaches also fill-in data for black objects (► Fig. 54.6). If no readings are obtained along a range of vertical angles in a specific direction, the space must be occupied by a black obstacle. Otherwise the rays would have hit some obstacle or the ground.

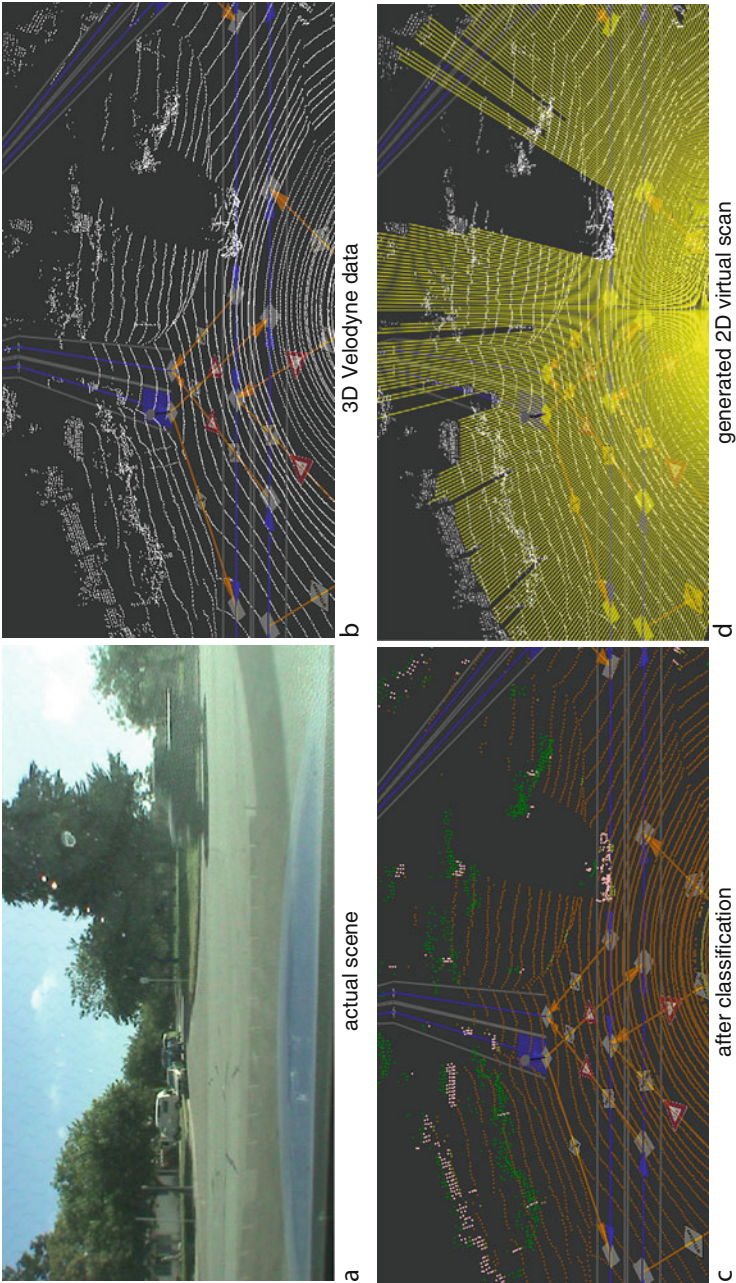
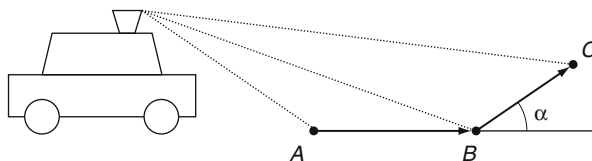


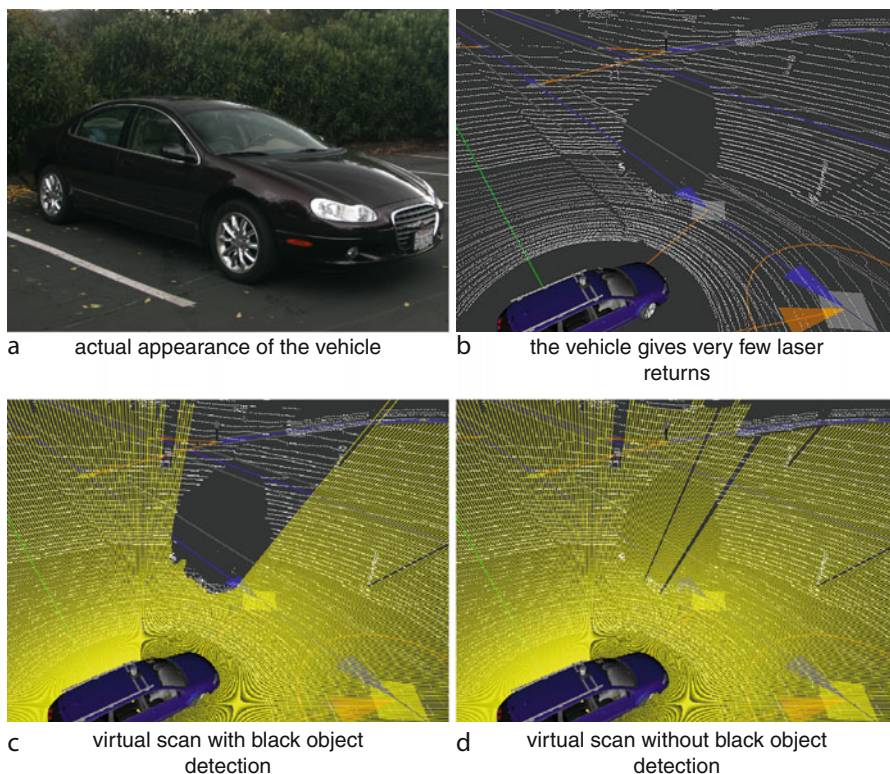
Fig. 54.4

An example of 3D to 2D conversion for Velodyne laser data. In (c) Velodyne data are colored by type: orange- ground, yellow- low obstacle, pink- medium obstacle, green- high obstacle. Only medium obstacles are projected to the resulting 2D scan. The figure is best viewed in color (available online). In greyscale, the brightest points are medium obstacles. In (d) white lines denote the resulting 2D virtual scan. Note the truck crossing the intersection, the cars parked on a side of the road and the white van parked on a driveway. On the virtual scan, all of these vehicles are clearly marked as obstacles, but ground, curbs, and tree tops are ignored. Images are extracted from an article by Petrovskaya and Thrun (2009)



■ Fig. 54.5

Ground readings can be determined by comparing angles between consecutive readings. If A , B , C are ground readings, then α is close to 0 and thus $\cos \alpha$ is close to 1. The figure is from Petrovskaya (2011)



■ Fig. 54.6

Detecting black vehicles in 3D range scans. White points represent raw Velodyne data. *Straight lines* in (c) and (d) represent the generated virtual scans. The car shape in (b), (c), and (d) denotes the position of the robot. Images are extracted from an article by Petrovskaya and Thrun 2009

For vision sensors, it is common to convert the image to gray scale, apply a Gaussian blur, and reduce the image resolution. When using stereovision, the pixels of the two images are matched. Then, a distance measurement is obtained through triangulation. With this technique, a pair of images can be transformed into a set of 3D points, represented in the *sensor*

coordinate system or in another space like the *disparity space* for some specific configuration of the sensor (Labayrade et al. 2002). Using successive images in an optical flow approach is another way to retrieve 3D data from a vision sensor.

2.2.4 Dynamics Models

Dynamics model describes motion of an object in probabilistic terms. Given a prior pose and velocity, the model defines a probability distribution over resulting poses at the next time step. In DATMO, it is common to assume a *constant velocity model*, in which the velocity of each tracked target stays constant for the duration of each time interval from $t-1$ to t . At each time step t , the velocity instantaneously evolves via addition of random noise based on maximum allowed acceleration a_{\max} and the time delay Δt from the previous time step $t-1$. More specifically, Δv is either sampled from a normal distribution $\mathcal{N}(0, a_{\max}\Delta t)$ or uniformly from $[-a_{\max}\Delta t, a_{\max}\Delta t]$.

Brownian Motion Model: The simplest model of motion is *Brownian motion*. This model is often used for pedestrians because people can rapidly change velocity and direction of motion. In this model, the pose evolves via addition of zero-mean Gaussian noise. The variance of the noise grows with Δt .

Linear Motion Model: Vehicles have more predictable motion than pedestrians due to higher inertia and non-holonomic constraints. Since the exact dynamics of tracked vehicles are unknown, it is common to use the *linear motion* model. In this case, the motion consists of perturbing orientation by $\Delta\theta_1$, then moving forward according to the current velocity by $v_t\Delta t$, and making a final adjustment to orientation by $\Delta\theta_2$. Given a maximum allowed turning rate $d\theta_{\max}$, $\Delta\theta_1$ and $\Delta\theta_2$ are sampled from a normal distribution $\mathcal{N}(0, d\theta_{\max}\Delta t)$ or uniformly from $[-d\theta_{\max}\Delta t, d\theta_{\max}\Delta t]$.

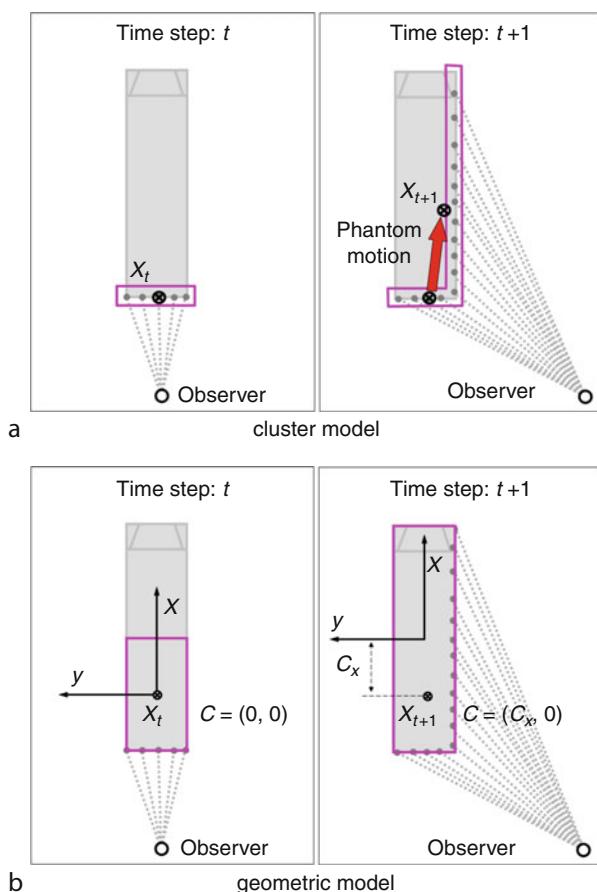
Bicycle Motion Model: A more refined motion model for vehicles is the *bicycle model*. This model uses angular velocity ω_t in addition to linear velocity v_t . As before, constant velocity model is assumed for both angular and linear velocities. The object moves along an arc, which is determined based on the linear and angular velocities. The equations of motion are more complex in this case. See Thrun et al. (2005), Sect. 5.4 for details.

2.3 Scene Models

2.3.1 Cluster-Based Models

One of the simplest and most common object models are cluster-based representations. In this case, tracked objects are represented as clusters of observed data points or features. Target's pose X_t represents the position of the geometric mean of the cluster in the world coordinate frame. The main drawback of this representation is that it is viewpoint dependent because the geometric mean does not account for unobserved parts of

the object. Once previously unobserved parts of the object come into view, the geometric mean shifts with respect to the object. This shift is seen as motion of the target because X_t moves with the geometric mean. This leads to incorrect estimates of the object's pose and velocity. For example, a robot can perceive a non-zero velocity for a stationary vehicle, simply because the robot moves around it and observes it from a different vantage point as [Fig. 54.7a](#) illustrates. This problem can be addressed using geometric object models, which are described below.



■ Fig. 54.7

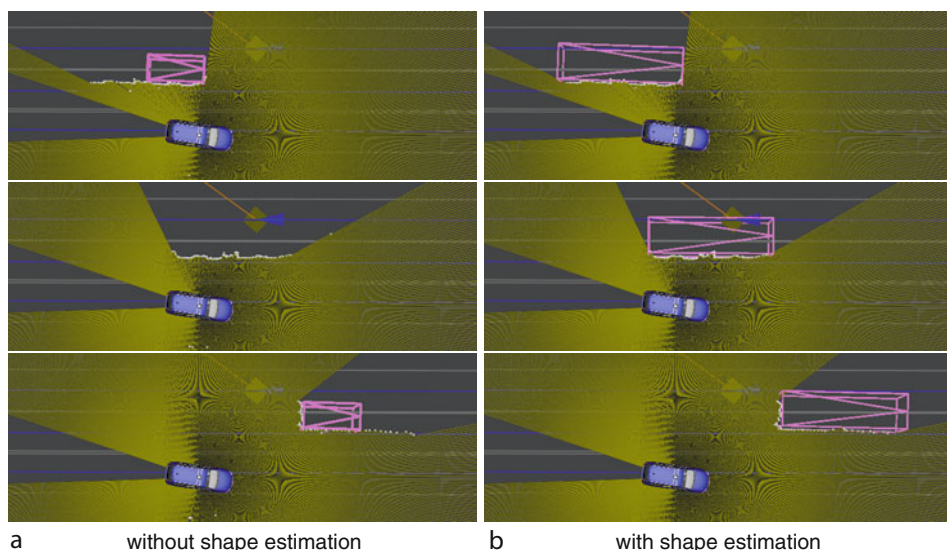
The effect of vantage point shift on estimated motion of the object. As the robot moves to observe a different side of a stationary bus, the geometric mean of observed points shifts with respect to the bus. (a) With cluster models, X_t is tied to the geometric mean of the points. Hence, a change in vantage point leads to phantom motion (*thick arrow*) of the object. (b) With geometric models, anchor point coordinates $C = (C_x, C_y)$ compensate for geometric mean shift, so that X_t remains stationary in world coordinates. The figure is from Petrovskaya (2011)

2.3.2 Geometric Models

As an alternative to clusters, objects can be represented by geometric shapes. Rectangular boxes are most common, although circles and ovals have also been used. Since one size and shape does not fit all targets well, some methods use several predefined shapes: one per class of objects. For example, pedestrians could be modeled as circles, bicycles as ovals, cars as rectangles, and buses as long rectangles. In this case, object class variable is added to the state for each target.

To obtain an even more accurate representation of objects, some approaches parameterize the shape and infer the shape parameters. For example, objects could be modeled as rectangles, for which width and length are added to the state variables. ➤ [Figure 54.8](#) illustrates the impact of shape inference on a real example.

When unobserved parts of an object come into view, with a geometric model, it is still possible to experience geometric mean shifts just as for the cluster-based representations discussed above. To overcome this problem, let X_t be the pose of an *anchor point* which is fixed somewhere on the object. Although the anchor point remains fixed on the object, the robot's belief of the anchor point position with respect to the object's center can change over time (➤ [Fig. 54.7b](#)). Initially, the robot sets the anchor point to be the center of what it believes to be the object's shape and thus, anchor point coordinates in the object's



■ Fig. 54.8

Shape inference on the example of a passing bus. Without shape estimation (a) the tracking results are poor because the geometric model does not fit the data well. Not only is the velocity estimated incorrectly, but the track is lost entirely when the bus is passing. With shape estimation (b) the bus is tracked successfully and the velocity is properly estimated. The car shape denotes the position of the robot. Images are from Petrovskaya and Thrun 2009

local coordinate system are $C = (0,0)$. Assume that the object's local coordinate system is tied to its center with the x -axis pointing directly forward. As the robot revises its knowledge of the object's shape, the local coordinates of the anchor point need to be revised accordingly to $C = (C_x, C_y)$. Thus, C_x and C_y are added to the target's state variables.

2.3.3 Grid-Based Models

Occupancy grids have been utilized to represent the scene for some time (Moravec 1988). Occupancy grids divide the scene into a grid of cells. The cells may be uniform or varying in size. Typically each cell has a value $P(O_i)$, representing the probability that something is occupying cell i . $P(O_i) = 0$ indicates a cell that is certainly empty, $P(O_i) = 1$ indicates that it is certainly occupied, and $P(O_i) = 0.5$ is typically used to represent cells for which no information is available.

In the intelligent vehicle context, the grid is typically used to represent an area in front of the vehicle. That is to say that rather than the grid being used to represent the complete environment at a large scale, the grid is essentially attached to the vehicle, in a robot-based frame, as described in [Sect. 1.2.1](#). This is because the grid is being used to model the scene and identify potential dangers, rather than building a global map. Work has been done using this moving grid along with local-SLAM to differentiate between moving and static obstacles.

In defining the grid, tradeoffs are made between accuracy and performance in specifying the grid size and cell size. Specifically, smaller cells allow for a more accurate representation of the scene. Similarly, a larger grid allows representation of more of the environment. However, additional cells also require additional computing resources, in terms of both memory for storing the grid, and computing power for updating the grids. The computing resources available and application-specific details (sensor range, sensor accuracy, accuracy requirements, etc.) should be used to define the grid.

In the grid-based approaches, concepts such as *objects* or *tracks* do not exist; they are replaced by other properties such as occupancy or elevation, which are directly estimated for each cell of the grid using both sensor observations and some prior knowledge. It might seem strange to have no object representations when objects obviously exist in real life environments. However, an object-based representation is not required for all applications. Where object-based representations are not pertinent, it is more useful to work with a more descriptive, richer sensory representation rather than constructing object-based representations with their complications in data association. For example, to calculate the risk of collision for a mobile robot, the only properties required are the probability distribution on occupancy and velocities for each cell in the grid. Variables such as the number of objects are inconsequential in this respect.

Occupancy grids are especially useful to fuse information from several sensors. In standard methods for sensor fusion in tracking applications, the problem of track-to-track association arises where each sensor contains its own local information. Under the standard tracking framework with multiple sensors, the problem of data association will

be further complicated: As well as the data association between two consecutive time instances from the same sensor, the association of tracks (or targets) between the different sensors must be taken into account as well.

In contrast, the grid-based approaches will not encounter such a problem. A grid-based representation provides a conducive framework for performing sensor fusion (Moravec 1988). Different sensor models can be specified to match the different characteristics of the different sensors, facilitating efficient fusion in the grids. The absence of an object-based representation allows easier fusing of low level descriptive sensory information onto the grids without requiring data association.

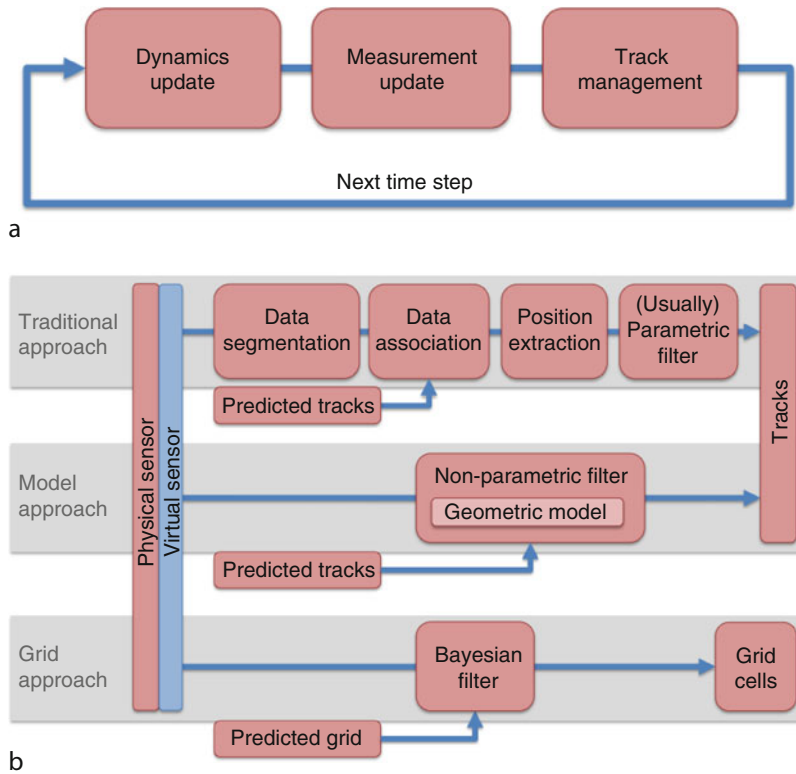
Grid mapping from sensor data: Occupancy grids are generally computed from the data provided by range finders. The classical approach is based on the *independent beam model* taking into account both *positive* and *negative* information, see ● Sect. 2.2.1.

Grid mapping is somewhat less common with vision sensors. With a single camera, integration over time is necessary to retrieve distance measurements, leading to approaches like visual-SLAM. Stereovision can provide 3D data, so it is current practice to use it as a distance sensor, considering a metric point cloud (Brailon et al. 2006). Stereospecific methods also exist. In Matthies and Elfes (1988) or Murray and Little (2000), the authors consider the first detected object for each column and suppose that it occludes the field ahead. This is obtained by finding the highest disparity value for each column of the image. The result is a ternary grid (free/occupied/unknown) and the approach is sensitive to outliers in the disparity map. An improvement has been proposed in Badino et al. (2007), where the authors propose to build such a grid in the u-disparity plan. Occupancy is then robustly estimated using dynamic programming. The result is a ternary grid, specifically designed for free field estimation. An improvement is proposed in Perrollaz et al. (2010b), where the authors propose a visibility-based approach to estimate the occupancy probability of each cell of the grid, in the u-disparity plane. This is done by estimating for each cell the probabilities of “being visible,” “containing an object,” and of “containing the road surface.”

3 Inference: Filtering and Tracking

In a Bayesian approach, the evolving state of a dynamic system is tracked using a *Bayesian filter*, which computes the current belief bel_t recursively from a prior belief bel_{t-1} using the Bayesian recursion equation (● 54.1). The filter consists of a *dynamics update*, which captures the evolution of the state between time steps, and a *measurement update*, which incorporates the most recent sensor measurements (Gordon 1993; Diard et al. 2003). In DATMO, separate Bayesian filters are used to track the state of each moving object (or cell). Updates of the Bayesian filter are followed by a track management phase, during which tracks are created and deleted and dependencies between targets are enforced. The resulting pipeline for DATMO is summarized in ● Fig. 54.9a.

There are two types of Bayesian filters: parametric and non-parametric (for an overview see Arulampalam et al. 2002). *Parametric filters* include variants of *Kalman filters* (KF) and represent the belief by a parametric function, a *Gaussian*. Since the actual belief is



■ Fig. 54.9

(a) Typical DATMO pipeline. (b) Three alternative approaches to the measurement update step in the DATMO pipeline. Each approach starts with sensor data on the left and produces tracks of moving objects or grids on the right. Predicted tracks and predicted grid are outputs of the dynamics update. The virtual sensor box is optional, but the rest are mandatory for each branch

usually non-Gaussian, approximations have to be made as in the extended Kalman filter (EKF) or the unscented Kalman filter (UKF). *Non-parametric filters* include *particle filters* (PF) and *histogram filters* (HF). They represent the belief by a set of points, which are positioned randomly in case of the PF and deterministically in HF. In PF, these points are called *particles*; in HF, they are *grid cells*.

Parametric methods have the advantage of being more efficient as their computational cost is polynomial in the dimensionality of the state. However, parametric methods are unable to represent complex beliefs, which often arise due to ambiguities in the data. They also often perform poorly if a sufficiently good estimate of the state is not available *a priori*. In contrast, non-parametric methods have the advantage of being able to represent an arbitrary belief even if it has multiple high probability regions (called *modes*). However, their computational cost is exponential in the dimensionality of the state and the base of the exponent tends to be large. This property is often referred to as the *curse of dimensionality* (MacKay 1998).

Several types of DATMO approaches have evolved to build around these advantages and disadvantages. To help organize the discussion in this chapter, these approaches can be roughly divided into three main branches (► [Fig. 54.9b](#)): the traditional approach, the model-based approach, and the grid-based approach. The first two branches represent two schools of thought for producing object level tracks: (i) build a pseudo-sensor of target positions by augmenting the physical sensor with a set of algorithms and then use a simple (usually parametric) filter (top branch), or (ii) use the raw sensor with a physical sensor model and employ a more advanced non-parametric filter (middle branch). In (i) most of the problem is solved while producing pseudo-measurements of target positions, whereas in (ii) most of the problem is taken care of by the inference method. Both of these approaches are discussed below in ► [Sects. 3.2](#) and ► [3.3](#).

The third approach (bottom branch in ► [Fig. 54.9b](#)) has a different goal than the first two. Instead of object level tracks of targets, it produces tracks of grid cells – a lower level representation of moving obstacles, which can be used directly for low level trajectory planning. This approach is described in ► [Sect. 3.4](#). To obtain object level tracks, the grid-based approach can be followed by a process similar to the traditional branch, but replacing the physical sensor with the grid representation, which is built by the grid-based approach. This method is described in ► [Sect. 3.4.3](#).

3.1 Track Management and Dynamics Update

3.1.1 Imposing Track Constraints

Although during inference, inter-track dependencies are ignored as was already mentioned in ► [Sect. 1.2.2](#), these dependencies clearly exist. For example, no two vehicles can occupy the same space at the same time. These dependencies are enforced during the track management phase by imposing constraints. The constraints may include assumptions that targets do not overlap, are surrounded by some amount of free space, and are located in *regions of interest* (ROI, e.g., on or near roads).

Tracks violating constraints can be deleted right away, can have their existence score decreased (as discussed below), or can be merged with other tracks in the case of overlapping targets.

3.1.2 Track Existence

Creation and deletion of tracks relies on *track existence* probability, which can be derived using the Bayes rule. This probability is usually kept in log form as an *existence score* which represents the *log-likelihood ratio* (LLR) first proposed by Sittler (1964) and summarized by Blackman et al. (2004). This approach is an application of the classical *sequential probability ratio test* (SPRT).

The existence score takes into account whether or not the target has been observed, how well it matched the observations, how well it fits the dynamics model, as well as expected probability of detection and expected density of false alarms. It can also include terms for *signal-to-noise ratio* (SNR). In short, if a target is observed and matches the data well then the score is increased, otherwise it is decreased.

3.1.3 Track Creation

After a measurement update, the sensor data not well explained by existing tracks are examined to initialize new tracks. This process can take several frames of data into account (e.g., from time steps $t-2$, $t-1$, and t) by running a Bayesian filter over these frames of data for a prospective target. For each frame of data, the existence score is computed for the prospective target. If the score is sufficiently high in all frames, the prospective target may be added to the list of tracked targets.

Note that detection of new targets is a more challenging problem than tracking of existing targets. Since the initial pose uncertainty for each new target is very large, large areas of space need to be examined for possible existence of new targets. In contrast, during tracking of existing targets, a prior pose estimate is already available from the previous time step and hence the pose uncertainty of each target is lower.

A number of techniques can be used to make detection of new targets more efficient. For example, it is common to restrict areas in which new targets can enter the scene. These areas may be defined by boundaries of the observed space (i.e., image boundaries for cameras and point cloud boundaries for 3D sensors) and by features of the environment (e.g., entrance lanes on freeways or intersections). Another approach is to perform some fast preprocessing of the data to determine locations, in which new targets may have appeared. For example, sufficiently large change has to occur in an area if it contains a moving object.

3.1.4 Track Deletion

Like for track creation, existence score can be used to determine if a particular track should be discontinued. The existence score is computed for each target at each time step. Once the score falls below a certain threshold, the track is deleted.

3.1.5 Dynamics Update

Once track management phase completes, all surviving tracks are propagated to the next time step. The propagation is performed using the dynamics update of the Bayesian filter used to track the targets. The dynamics update relies on probabilistic laws of target

dynamics, which we discussed in 🔗 Sect. 2.2.4. In effect, the dynamics update computes the prediction distribution for each target $\overline{bel}_t := p(S_t | Z_1, \dots, Z_{t-1})$, which is given by the integral in the Bayesian recursion equation (🔗 54.1).

3.2 Traditional DATMO

The traditional pipeline of DATMO corresponds to the top branch in 🔗 Fig. 54.9b. It usually relies on variants of Kalman filters, although some recent approaches have used particle filters. The distinguishing characteristic of the traditional approach is that most of the work is done prior to application of the filter. The data are segmented into meaningful pieces using clustering or pattern recognition methods as described in 🔗 Sects. 3.2.1 and 🔗 4.1, respectively. The pieces of data are associated with targets using data association methods described below in 🔗 Sect. 3.2.2. This stage can be challenging because of the association ambiguities that often arise. Finally, for each target, a pseudo-measurement of position is produced by taking geometric mean of the data assigned to each target. This position estimate is then used in a Kalman filter variant (or a particle filter) to update the belief of each target's state.

A broad spectrum of DATMO approaches fall into the traditional approach category. While early work has used simple clustering techniques to segment data (🔗 Sect. 3.2.1), recent work within the traditional branch partitions data based on advanced pattern recognition techniques to improve detection of targets (🔗 Sect. 4.1).

3.2.1 Data Segmentation

Among the simplest methods for data segmentation are the *clustering methods*. These methods can be applied directly to range data. The data points are segmented into clusters based on range discontinuities. These clusters then need to be assigned to targets during the data association phase; although multiple clusters may need to be associated with the same target.

A slightly more complex approach is to look for specific features in the data. Straight lines and letter L's are common features for range data. Harris corners and edges are common for vision data. Again these features need to be associated to targets with possibly multiple features assigned to each target.

More sophisticated techniques for data segmentation rely on pattern recognition techniques, which are discussed in 🔗 Sect. 4.1.

3.2.2 Data Association

Once the data have been segmented into pieces, these pieces need to be associated to targets. Below is a summary of data association methods. See Cox (1993) for a detailed review.

One of the simplest methods for data association is the *nearest neighbor* (NN) method, which assigns each piece of data to the closest predicted target. This method is widely used when update rates are high enough to make this assignment unambiguous. A more sophisticated method is the *global nearest neighbor* (GNN), which ensures that each piece of data is assigned to one target. It is useful for recognition-based methods, which ensure that each piece of data represents the entire tracked object (e.g., pedestrian or vehicle) and hence a one-to-one correspondence between targets and pieces of data is appropriate.

More advanced methods have been developed for situations where ambiguities are plentiful. The two most common approaches are the *multiple hypothesis tracking* (MHT) algorithm and the *joint probabilistic data association filter* (JPDAF).

MHT Algorithm: Originally developed by Reid (1979), this algorithm maintains multiple association hypotheses between targets and measurements. The MHT framework can handle situations where the measurements arise from a varying number of targets or from background clutter. For example, in Fig. 54.10, there is a single hypothesis H_{t-1}^1 at time $t-1$ containing targets T_{t-1}^1 and T_{t-1}^2 . At time t , the new observations are segmented into three pieces denoted by Z_t^1 , Z_t^2 , and Z_t^3 . The new possible hypotheses are formed from the prior step hypothesis by associating the new measurements to the existing targets or by initializing new targets from each new measurement. Thus, these associations give rise to a number of hypotheses at time t , of which only two are shown in the figure.

For each of the obtained hypotheses, a hypothesis score is computed by summing the track existence scores (from Sect. 3.1.2) of all targets within it. The hypothesis probability can then be computed from the hypothesis score.

In practical situations, several issues arise from the application of this method. The most serious is the combinatorial increase in the number of generated tracks and

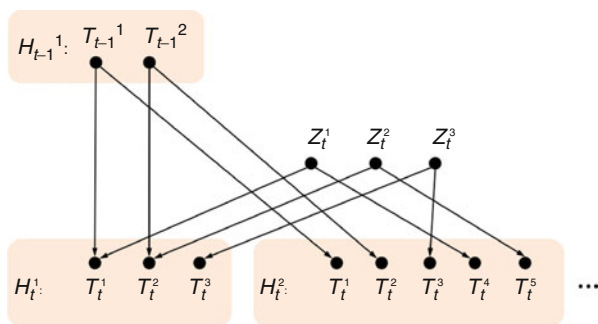


Fig. 54.10

Example of a data association problem with two targets T_{t-1}^1, T_{t-1}^2 from time $t-1$ and three new measurements Z_t^1, Z_t^2, Z_t^3 . A single hypothesis H_{t-1}^1 from the prior step splits into multiple hypotheses at time t , of which only H_t^1 and H_t^2 are shown. Arrows show the associations made

hypotheses. Several techniques such as tracks clustering or pruning can be used to avoid this increase. For further information on MHT and its variants, see Blackman et al. (2004).

JPDFAF Algorithm: The algorithm was proposed by Fortmann et al. (1983) based on the *probabilistic data association* concept introduced by Bar-Shalom and Jaffer (1972). Unlike the MHT algorithm, JPDAF does not suffer from the combinatorial explosion. It is similar to GNN in that a single association hypothesis is maintained. However, unlike GNN, JPDAF does not make a hard assignment of measurements to targets. Instead, it makes a soft assignment by considering the probability of each measurement being assigned to each target.





More specifically, suppose we have K targets and M_t data segments $Z_t^1, \dots, Z_t^{M_t}$. Let the probability of measurement m being caused by target k be denoted by β_{mk} . Then for target k the measurement update is carried out considering all possible assignments of data segments:

$$p(Z_t|S_t^k) = \eta \sum_{m=1}^{M_t} \beta_{mk} p(Z_t^m|S_t^k), \quad (54.2)$$

where η is a normalization constant. For details on how to approximate the probabilities β_{mk} see Schulz et al. (2003).

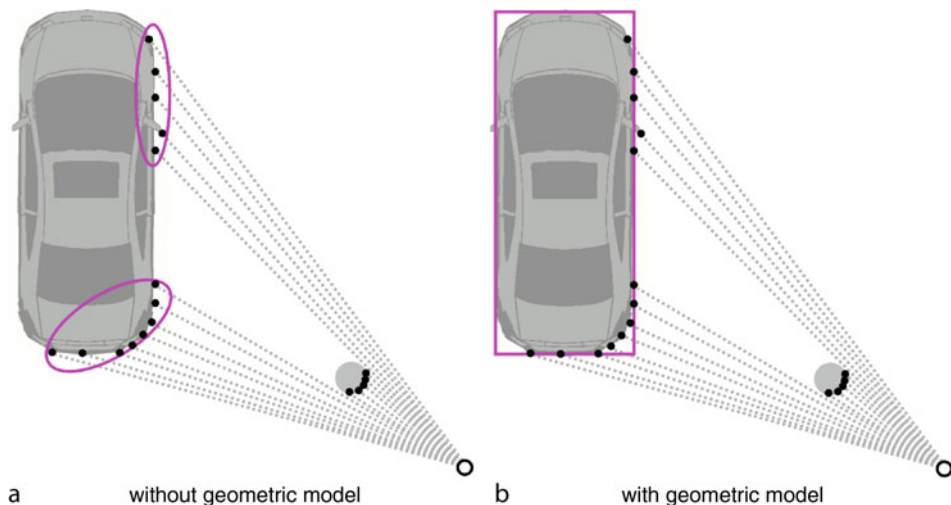
A number of extensions and variants of JPDAF have been developed (Cox 1993). Most methods utilizing JPDAF rely on parametric belief representations (i.e., Kalman filters). Such methods are widely used in DATMO approaches. Non-parametric variants of JPDAF have also been developed (Schulz et al. 2003; Vermaak et al. 2005), although these methods have yet to find their way into DATMO for autonomous driving.

3.3 Model-Based DATMO

The model-based approach corresponds to the middle branch in  Fig. 54.9b. It works by directly modeling the physical sensor and the moving objects using a physical sensor model ( Sect. 2.2.1) and geometric models of objects ( Sect. 2.3.2). Most of the work here is done by the filter itself. Separate data segmentation and association steps are not required because the geometric object model helps associate data points to targets (see  Fig. 54.11). The main challenge is to make the filter efficient enough to meet the high demands of autonomous driving.

3.3.1 Rao-Blackwellization

Inference efficiency can be improved with a hybrid representation of the belief using a *Rao-Blackwellized particle filter* (RBPF). The concept of Rao-Blackwellization dates back to Rao (1945) and Blackwell (1947) and has been used with great success in SLAM (Murphy and Russell 2001; Montemerlo 2003). In RBPF, the belief over some parameters



■ Fig. 54.11

Scans from objects are often split up into separate clusters due to occlusion (a). Geometric model helps interpret the data properly (b). Ovals and rectangle group together points that have been associated together. In (b), the rectangle also denotes the geometric model. Gray car shape and filled circle are objects. Gray dotted lines represent laser rays. Black dots denote laser data points. The figure is from Petrovskaya (2011)

of the state is represented by particles with each particle maintaining a Gaussian belief over the remaining parameters. This technique effectively reduces the dimensionality of the state from the perspective of the particle filter.

Let Ω denote the vector of object shape parameters (e.g., $\Omega = (W, L, C_x, C_y)$). Then the full belief is $bel_t := p(X_{1:t}, v_{1:t}, \Omega | Z_{1:t})$, where $1:t$ is a shorthand for $1, \dots, t$. For Rao-Blackwellization, the belief is split into two conditional factors:

$$bel_t = p(X_{1:t}, v_{1:t} | Z_{1:t}) p(\Omega | X_{1:t}, v_{1:t}, Z_{1:t}). \quad (54.3)$$

The first factor represents the belief about the object motion, whereas the second factor represents the belief about the object shape, conditioned on its motion. Let A_t denote the first factor and B_t denote the second factor. The motion belief is represented by particles and the shape belief is represented in Gaussian form. The dynamics update of RBPF is the same as standard particle filter (PF): The particles are resampled and propagated according to the dynamics model. Like for PF, the measurement update involves computation of the importance weights. For RBPF, the correct importance weights can be shown to be :

$$w_t = \frac{p(X_{1:t}, v_{1:t} | Z_{1:t})}{p(X_{1:t}, v_{1:t} | Z_{1:t-1})} = \mathbf{E}_{B_{t-1}}[p(Z_t | \Omega, X_t)]. \quad (54.4)$$

In words, the importance weights are the expected value of the measurement likelihood with respect to the object shape prior. Using Gaussian approximations of

B_{t-1} and the measurement model, this expectation can be computed in closed form. See Petrovskaya (2011) for derivations.

3.3.2 Scaling Series Algorithm

Efficiency can also be improved using an iterative annealing method, such as a *Scaling Series particle filter* (SSPF). SSPF gradually refines the belief from very coarse to fine resolution and simultaneously adjusts the annealing temperature. At each iteration, it uses a divide-and-conquer strategy and thus, it can be exponentially faster than the basic particle filter. Even though in principle, SSPF cannot escape the curse of dimensionality, it effectively reduces the base of the exponent (e.g., from 30 down to 6). Hence, the curse's effect is greatly diminished allowing the approach to outperform PF by several orders of magnitude. Detailed description of SSPF can be found in Petrovskaya and Khatib (2011).

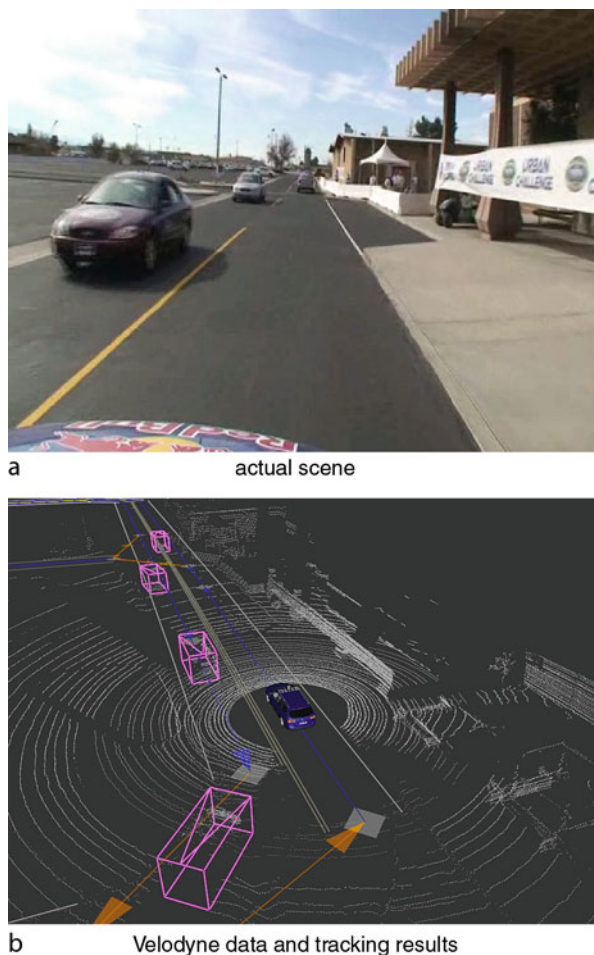
An example combining the use of SSPF and RBPF techniques to solve the DATMO problem can be found in Petrovskaya and Thrun 2009, where the resulting approach tracks vehicles in real time (see ● Fig. 54.12). Since the model-based approach leads to accurate estimation of motion, it is possible to build full 3D models of passing vehicles (● Fig. 54.13) by collecting all scan points assigned to each target (Petrovskaya 2011). The 3D models can in turn be used for even more accurate motion estimation, which is particularly useful for slow-moving vehicles near intersections.

3.3.3 MCMC Approach

The DATMO problem can be posed as a batch problem over a sliding window of a fixed number of frames. In this form, it can be solved using the *Markov Chain Monte Carlo* (MCMC) approach. See Andrieu et al. (2003) for an introduction to MCMC and related concepts.

MCMC methods produce samples from a desired probability distribution by constructing a *Markov chain* for which the *equilibrium distribution* is equal to the desired distribution. When these methods “walk” the chain for a sufficiently large number of steps, the resulting state of the chain can be used as a sample from the desired distribution. While it is usually easy to construct a Markov chain with the desired distribution as its equilibrium, determining the required number of steps (known as the *mixing time*) can be difficult.

The *Metropolis-Hastings* version of MCMC was proposed by Metropolis et al. (1953) and Hastings (1970). It uses a proposal distribution and accepts or rejects the next state based on a ratio test, which takes into account both the desired distribution and the proposal distribution. Although this algorithm will eventually converge for any proposal distribution, a more informed proposal distribution greatly improves efficiency of the approach. For this reason, Vu and Aycard (2009) used a *data driven* MCMC technique and generated object hypotheses with a detection module, which identifies moving parts of dynamic objects.



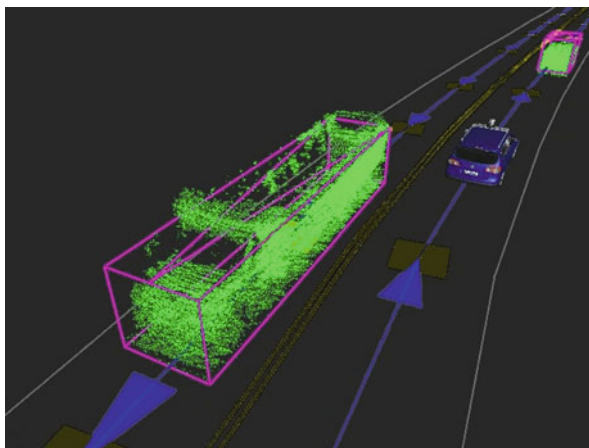
■ Fig. 54.12

Tracking results on course A at the 2007 Urban Grand Challenge. The 3D boxes in (b) denote the tracked vehicles, the white points represent Velodyne data, the car shape in the center shows the position of the robot. Images are from Petrovskaya and Thrun (2009) and Petrovskaya (2011)

Further, their approach uses discrete models of objects: rectangles of several fixed sizes for cars, buses, and bikes, and points for pedestrians. Each class of objects also has its own dynamics model, which is integrated into the framework using an *interacting multiple model* (IMM) filter (Vu 2009).

3.4 Grid-Based DATMO

The grid-based DATMO approach corresponds to the bottom branch in Fig. 54.9b. The Bayesian filtering paradigm can be applied for estimation and filtering in the occupancy



■ Fig. 54.13

3D shape inference for vehicle tracking. The *green points* (or the bright grey points in greyscale version) show the accumulated Velodyne points assigned to each vehicle over multiple frames. Tracking with anchor point coordinates allows us to align the points from different frames on per-vehicle basis. Blue car (or dark gray car in grayscale version) shows the position of the ego-vehicle. Image is from Petrovskaya (2011)

grid framework. This provides multiple advantages compared to static grids estimated independently for each frame of sensor data:

- It allows estimation of the velocity of each cell of the grid, hence modeling the dynamic environment.
- Since it takes into consideration successive measurements over time, it allows retention of information about occupancy for regions of the scene that are occluded.
- Through filtering, it can remove errors that are present on a unique frame of measurements.

Moreover, in the classical methodology presented in the top branch of Fig. 54.9b, the problems of data association and state estimation are highly coupled, and an error in either component leads to erroneous outputs. The grid filtering methodology makes it possible to decompose this highly coupled relationship by avoiding the data association problem, in the sense that the data association is handled at a higher level of abstraction. A few methods have been proposed in the literature in order to adapt Bayesian filtering to occupancy grids, the first and more popular approach is the *Bayesian Occupancy Filter (BOF)*, which is briefly described below.

3.4.1 The Bayesian Occupancy Filter

The *Bayesian Occupancy Filter (BOF)* addresses the dynamics of the environment using a two-step mechanism, derived from the Bayesian filter. This mechanism estimates, at

each time step, the state of the occupancy grid by combining a prediction step (history) and an estimation step (incorporating new measurements). The consideration of sensor observation history enables robust estimation in changing environments (i.e., it allows processing of temporary objects, occlusions, and detection problems).

The state space of the BOF is a 2-dimensional grid. Each cell of the grid is associated with a probability of occupancy and a probability distribution on the velocity of the occupancy associated with the cell. Contrary to the initial formulation presented in Coué et al. (2006), this formulation does not allow for overlapping objects. On the other hand, it allows for inferring velocity distributions and reduces the computational complexity.

Define the following variables:

- C is an index that identifies each 2D cell of the grid.
- A is an index that identifies each possible antecedent of the cell c over all the cells in the 2D grid.
- $Z_t \in \mathcal{Z}$ where Z_t is the random variable of the sensor measurement relative to the cell c .
- $V \in \mathcal{V} = v_1, \dots, v_n$ where V is the random variable of the velocities for the cell c and its possible values are discretized into n cases.
- $O, O^{-1} \in \mathbf{O} \equiv \{occ, emp\}$ where O represents the random variable of the state of c being either “occupied” or “empty.” O^{-1} represents the random variable of the state of an antecedent cell of c through the possible motion through c . For a given velocity $v_k = (v_x, v_y)$ and a given time step Δt , it is possible to define an antecedent for $c = (x, y)$ as $c^{-k} = (x - v_x \Delta t, y - v_y \Delta t)$.

According to Bayes’ rule and dependency assumptions, the decomposition of the joint distribution of the relevant variables can be expressed as:

$$P(C, A, Z, O, O^{-1}, V) = P(A)P(V|A)P(C|V, A)P(O^{-1}|A)P(O|O^{-1})P(Z|O, V, C)$$

With this decomposition, each component can be associated with a semantic and a parametric form. Particularly, $P(O|O^{-1})$ is the conditional distribution over the occupancy of the current cell, which depends on the occupancy state of the previous cell. It is defined as a transition matrix:

$$T = \begin{bmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix},$$

which allows the system to use a constant velocity hypothesis as an approximation. In this matrix, ϵ is a parameter representing the probability that the object in c does not follow the null acceleration model, i.e., ϵ models the prediction error.

The aim of filtering in the BOF grid is to estimate the occupancy and grid velocity distributions for each cell of the grid, $P(O, V|Z, C)$. The two stages of prediction and estimation are performed for each iteration. In the context of the BOF, prediction propagates cell occupancy probabilities for each velocity and cell in the BOF ($P(O, V|C)$). During estimation, $P(O, V|C)$ is updated by taking into account its observation $P(Z|O, V, C)$ to obtain its final Bayesian filter estimation $P(O, V|Z, C)$. The result from the Bayesian filter estimation is then used for prediction in the next iteration.

3.4.2 BOF Implementation

When implementing the BOF, the set of possible velocities is discretized. One way of implementing the computation of the probability distribution is in the form of histograms. The following equations are based on the discrete case. Therefore, the global filtering equation can be obtained by:

$$P(V, O|Z, C) = \frac{\sum_{A, O^{-1}} P(C, A, Z, O, O^{-1}, V)}{\sum_{A, O, O^{-1}, V} P(C, A, Z, O, O^{-1}, V)}, \quad (54.5)$$

The global filtering equation (54.5) can actually be separated into three stages. The first stage computes the prediction of the probability measure for each occupancy and velocity:

$$\begin{aligned} \alpha(occ, c_k) &= \sum_{A, O^{-1}} P(A) P(v_k|A) P(C|V, A) P(O^{-1}|A) P(occ|O^{-1}), \\ \alpha(emp, v_k) &= \sum_{A, O^{-1}} P(A) P(v_k|A) P(C|V, A) P(O^{-1}|A) P(emp|O^{-1}) \end{aligned}$$

This is performed for each cell in the grid and for each velocity. Prediction for each cell is calculated by taking into account the velocity probability and occupancy probability of the set of antecedent cells.

With the prediction of the grid occupancy and its velocities, the second stage consists of multiplying by the observation sensor model, which gives the unnormalized Bayesian filter estimation on occupancy and velocity distribution:

$$\begin{aligned} \beta(occ, v_k) &= P(Z|occ, v_k) \alpha(occ, v_k), \\ \beta(emp, v_k) &= P(Z|emp, v_k) \alpha(emp, v_k). \end{aligned}$$

Similarly to the prediction stage, these equations are performed for each cell occupancy and each velocity. The marginalization over the occupancy values gives the likelihood of a certain velocity:

$$l(v_k) = \beta(occ, v_k) + \beta(emp, v_k).$$

Finally, the normalized Bayesian filter estimation on the probability of occupancy for a cell C with a velocity v_k is obtained by:

$$p(occ, v_k|Z, C) = \frac{\beta(occ, v_k)}{\sum_{v_k} l(v_k)}. \quad (54.6)$$

The occupancy distribution in a cell can be obtained by the marginalization over the velocities and the velocity distribution by the marginalization over the occupancy values:

$$P(O|Z, C) = \sum_V P(V, O|Z, C), \quad (54.7)$$

$$P(V|Z, C) = \sum_O P(V, O|Z, C). \quad (54.8)$$

3.4.3 Object Level Tracking

There are often times where object level representations are required. The philosophy of the BOF is to delay the problem of data association and as a result does not contain information with respect to objects. A natural approach to obtain an object level representation from the BOF grids is to introduce grid-based clustering to extract object hypothesis and an object level tracker to handle the hypothesis extracted.

Approaches similar to image processing can be advantageously used for clustering the grid, e.g., a search for connected components after thresholding. When using a dynamic grid estimation algorithm, like the Bayesian Occupancy Filter, it is possible to take advantage of the estimated velocities, in order to improve the clustering. For instance, two very close objects with different velocities would be in the same cluster without velocity estimation, while they could be separated.

After segmentation of the grid, clusters representing the different objects of the scene can be tracked over time. A classical algorithm, like the JPDAF presented in [Sect. 3.1](#), can be used for tracks management. However, in the cluttered environment with numerous moving objects, the JPDAF suffers from the combinatorial explosion of hypotheses. The *Fast Clustering and Tracking Algorithm (FCTA)* is more adapted to the BOF framework and to real-time processing, since it uses a ROI-based approach to facilitate the association stage (Mekhnacha et al. 2008).

FCTA could be roughly divided into a clustering module, an ambiguous association handling module, and a tracking and track management module.

Clustering: The clustering module takes the occupancy/velocity grid of the BOF as the input and extracts object level reports from it. A natural algorithm to achieve this is to connect the eight-neighbor cells according to an occupancy threshold. In addition to the occupancy values, a threshold of the Mahalanobis distance between the velocity distributions is also employed to distinguish the objects that are close to each other but with different moving velocities. In order to avoid searching for clusters in the whole grid, the predicted targets states are used as a form of feedback, by predicting regions of interest (ROI) in which the clustering process starts.

A report for the tracker is a 4-dimensional observation corresponding to the position and the velocity of an extracted cluster. The 2D position component of this vector is computed as the mass center of the region, and the 2D velocity component is the weighted mean of the estimated velocities of all cells of the cluster. These estimations come with a covariance matrix representing the uncertainty of the observation.

Re-clustering and track merging: The output of this clustering module leads to three possible cases (a) no observation, where the object is not observed in the ROI, (b) ambiguity free, where one and only one cluster is extracted and is implicitly associated with the given object, and (c) ambiguity, where the extracted cluster is associated with multiple objects. Therefore, a Bayesian algorithm is used at this stage for tracks splitting and merging.

Tracks management: After all the existing tracks are processed, the non-associated cells are processed to extract clusters as observations for potential new targets, using a region growing strategy from some “cluster seeds.” Classically in the FCTA, the prediction and estimation of the targets are accomplished by attaching a Kalman filter with each track, similar to the traditional DATMO approach. A confidence on the existence of each track is continuously estimated, thus allowing to delete tracks with low confidence.

An example of result using the complete grid-based DATMO approach is presented in

► [Fig. 54.14](#).

4 Pattern Recognition and Sensor Fusion

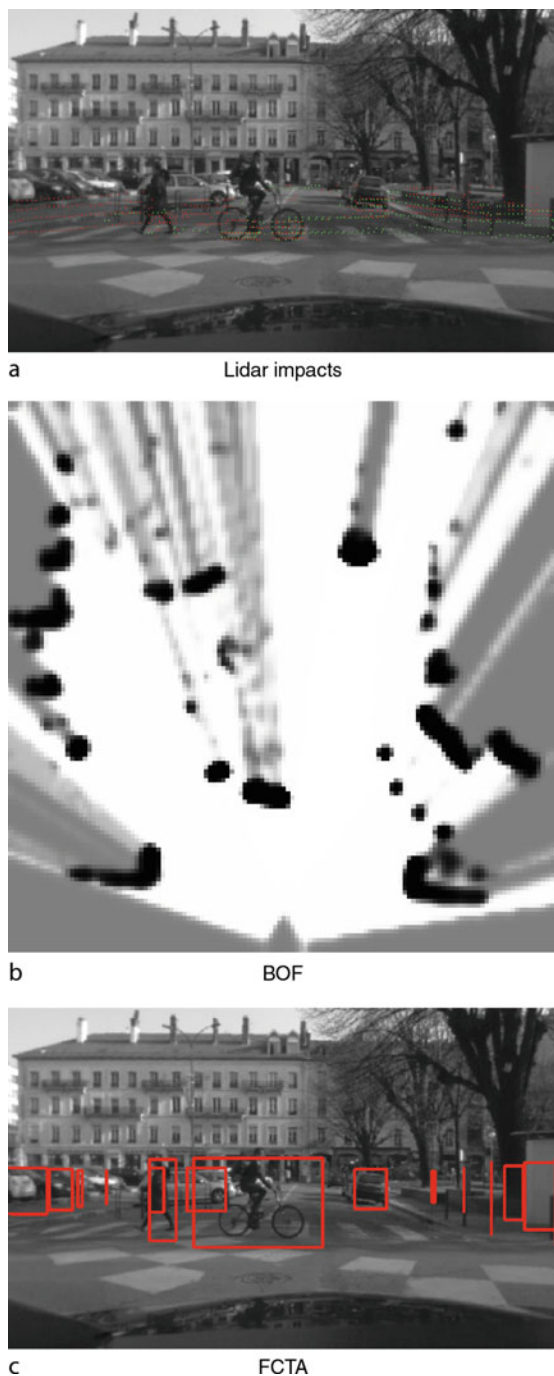
4.1 Pattern Recognition

4.1.1 Vision-Based approaches

Object and people detection from vision data has a long history and a large body of literature exists in this area. Most of the existing approaches can be classified into one of two major kinds of methods: window scrolling approaches and parts-based approaches (see ► [Fig. 54.15](#)).

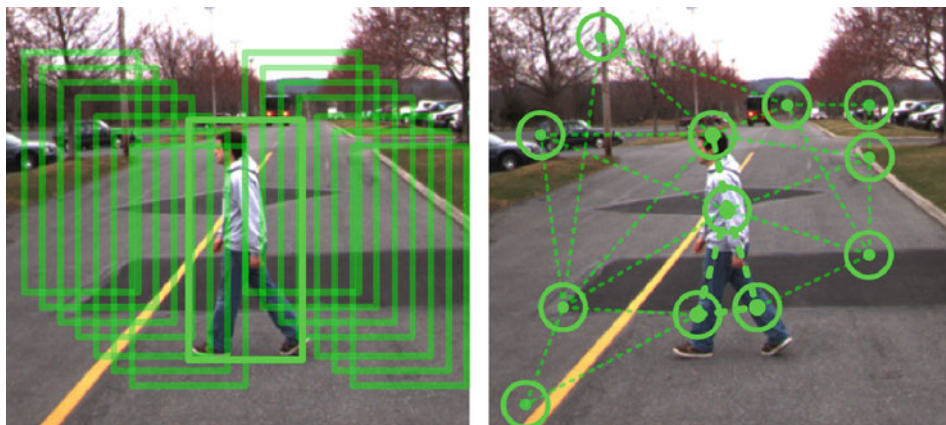
- Window scrolling approaches use a rectangular frame denoted as window that is shifted over a given image at all possible positions and scales to search for potential object occurrences. For training, a fixed-size window is used for computing features by employing a large hand-labeled dataset containing positive (people) and negative data (background). In a similar way, *silhouette matching* works by matching a contour of a person at different scales and positions using an edge-to-contour metric such as the Chamfer distance (Borgefors 1986). Seminal works in this area are Gavrila (2000) and Papageorgiou and Poggio (2000).
- Parts-based approaches use the fact that most objects as well as the human body consist of connected parts. Therefore, part appearances are learned and their geometrical relation is stored in a probabilistic way. This method produces detections when parts are represented in a geometrically consistent way in the image. Seminal works in this area are Viola et al. (2003), Mohan et al. (2001), and Mikolajczyk et al. (2004).

Recently, more mature methods that achieve high detection rates in crowded scenes have been presented (see Felzenszwalb et al. 2008; Tuzel et al. 2007; Dalal and Triggs 2005; Leibe et al. 2005; Wu and Nevatia 2005). Two datasets for people detection, captured from a moving vehicle, have been published (Enzweiler and Gavrila 2009; Dollar et al. 2009) in order to assess the performance of several detection algorithms. The conceptually simple Histogram-of-Oriented-Gradients detector by Dalal and Triggs (2005) performs very well, especially for detecting frontal views of people and people represented at intermediate sizes in the images. Haar-based detectors such as the one by Viola et al. (2003) are more suited in detecting smaller sizes. If pedestrians are represented at bigger scales and



■ Fig. 54.14

Example of lidar based BOF DATMO: (a) a urban scene, with dots representing the laser impacts, (b) the local occupancy grid estimated by BOF, (c) tracked objects extracted by FCTA



■ Fig. 54.15

The main approaches to object detection in computer vision. Window scrolling (*left*) and by-parts reasoning (*right*). The first classifies the image enclosed by the rectangle that has to be swept over the image at each position and at different scales. The latter classifies an object by reasoning on the geometrical disposition of object parts

if the aim is to also extract the articulation pose, then component-based methods are expected to perform better (see Leibe et al. 2005; Felzenszwalb et al. 2008). We note that many of these algorithms have been designed for people detection, but they also work for any other object category, for example cars.

The implicit shape model (ISM) algorithm by Leibe et al. (2005) is a generative detection method in which the detector is based on local features that independently cast votes for a common object center. The input for training is a set of images of people in which foreground and background is annotated via a binary mask. The first step of training is to collect all descriptors (SIFT, shape context, etc.) from interest points (Hessian-Laplace, Harris-Laplace, etc.). Then, a clustering algorithm is employed to reduce and generalize the numbers of descriptors present in the image set. The authors suggest using agglomerative clustering with average linkage, run with a fixed Euclidean distance θ . The resulting clustered centroids constitute the elements of the object *codebook*. Each element of the codebook is then matched to the training set images by using the same distance θ . The matched descriptors are used to note the relative position, with respect to the object center, of the associated interest points as a tag of the codebook element. To summarize, each element of the codebook is a generalized representation of an object patch and it is associated to a set of vectors indicating the positional displacement with respect to the object center.

During detection, a new image is acquired and interest points and associated descriptors are computed. Descriptors are matched with the codebook by using the distance θ . Matches are used to cast votes for object centers from the positions of the associated interest points. Votes are collected in a 3D voting space, defined by the x , y image coordinates and scale s . This procedure defines a generic Hough transform (GHT),

in which the modes of the votes distribution define object detection hypotheses. A clever form of mode estimation, sensitive to scale errors amplification, is represented by a modification of the standard mean-shift technique (see Comaniciu and Meer 2002), in which the kernel size is inflated anisotropically with respect to the scale (balloon kernel). In order to refine results in case of overlapping detection, a maximum descriptor length (MDL) cost is used that takes into account that a feature is not sharable and it must belong to one object or to another. A quadratic boolean programming problem is formulated in order to solve this best assignment. Furthermore, this technique can be used to generate objects segmentations by associating bitmap patches to successfully resolved feature assignments. One major disadvantage of ISM is that it relies on standard image features that have not been designed specifically for object detection. A solution to this drawback is provided by Felzenszwalb et al. (2008) that combines the advantages of ISM and HOG. It uses HOG-based classifiers to detect body parts and it assembles them in a probabilistic shape model.

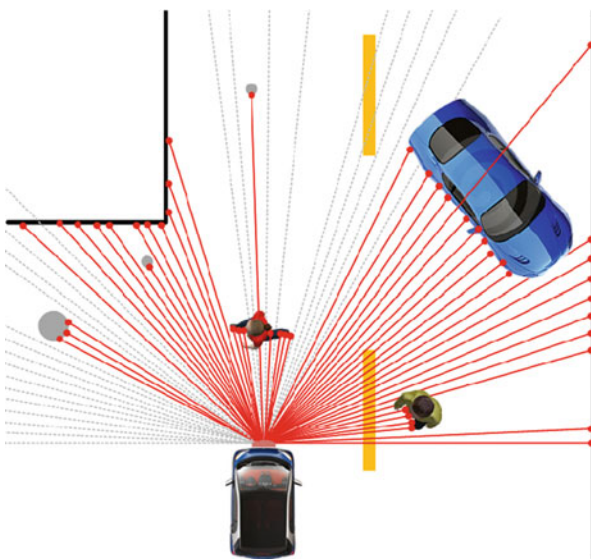
A more classic window-based approach is used by the HOG (histogram of oriented gradients) detector of Dalal and Triggs (2005). This detector is based on discriminative classification; thus, negative and positive image training sets are needed. Each training sample is specified as a rectangle enclosing the object. This rectangle (or window) has fixed size and it is used for computing features at fixed locations. Precisely, the window is tessellated in evenly overlapping square cells. A histogram of gradient directions is collected from the contents of each cell. Cells are organized into overlapping 2×2 blocks for computing a histogram normalization that takes into account larger area to encode robust illumination change. All the normalized block descriptors (HOG descriptors) are concatenated into a combined feature vector, which is used in a conventional Support Vector Machine (SVM) with a linear kernel. The negative and positive image training sets are used to train the SVM.

During detection, a new image is acquired and window is scrolled through the image at each position and scale. HOG is computed in each step and classified via the learned SVM. A non-maxima suppression is used in order to refine results and solve overlapping detections ambiguity. A computationally faster version of HOG, based on Adaboost SVM cascades on a non-regular window tessellation, has been developed by Zhu et al. (2006). Other variants leverage the parallel processing power of GPU (see Prisacariu and Reid 2009) achieving 120 Hz in classification speed. Occlusion handling and overall performance has been improved with the HOG-LBP detector, a novel detector that combines the HOG feature concept with local binary patterns classification (LBP) (see Wang et al. 2009). People detection at very small scales has been addressed by Spinello et al. (2009a), in which small edge segments, extracted via a superpixel segmentation, are assembled in a discriminative voting approach.

For car detection, simplistic computer vision approaches that exploit shadow detection (Dickmanns et al. 1994), trunk-frontal symmetries (Kuehnle 1991), rectangular shapes (Bertozzi et al. 1997), or vehicle lights (Cucchiara and Piccardi 1999) are nowadays overcome by HOG/ISM detectors trained with databases of car images (Ess et al. 2009; Leibe et al. 2007). Cars are learned with several viewpoint-dependent classifiers in order to account for the difference in appearance with respect to the object viewpoint.

4.1.2 Object and People Detection from 2D Laser Data

To detect objects in 2D laser range scans (● Fig. 54.16), several approaches have been presented in the past (see, e.g., Fod et al. 2002; Kleinhagenbrock et al. 2002; Schulz et al. 2003; Xavier et al. 2005; Arras et al. 2007). These approaches extract features from the 2D scan data, which are used to classify the data points into people or background. The features used can be motion features, geometric features such as circles and lines, or a combination of these, and they are either predetermined or can be learned from hand-labeled data as in Arras et al. (2007). However, most of these approaches have the disadvantage that they disregard the conditional dependence between data in a close neighborhood. In particular, they cannot model the fact that the label l_i of a given laser segment S_i is more likely to be l_j if we know that l_j is the label of S_j and that S_j and S_i are neighbors. One way to model this conditional dependency is to use Conditional Random Fields (CRFs) (Lafferty et al. 2001), as shown by Douillard et al. (2008). CRFs represent the conditional probability $p(\mathbf{l}|\mathbf{s})$ using an undirected cyclic graph, in which each node is associated with a hidden random variable l_i and an observation s_i . For example, \mathbf{l} can be a vector of discrete labels that range over the three different classes “pedestrian,” “car” and “background,” where \mathbf{s} are the feature vectors extracted from the 2D segments in the laser scan. To obtain the structure of the CRF, the 2D scan needs to be clustered first, which can



■ Fig. 54.16

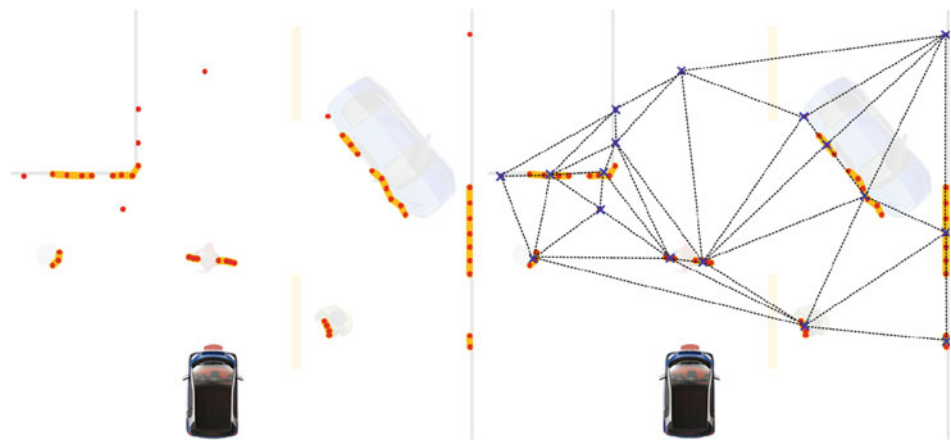
Example of a 2D laser scan. Laser beams are shown as *lines*, while *circles* represent the measured points. *Dotted lines* indicate out-of-range readings which, apart from the absence of an obstacle, can be caused by material reflections, sun light effects, and a too large incidence angle between the laser beam and the surface normal

be done using jump-distance clustering, a method that assigns two adjacent points to the same cluster if they are not further away from each other than a given threshold. As a result, all further reasoning is done on the clusters instead of the individual data points. Then, in a second step, the graph structure needs to be computed. This can be done using a Delaunay triangulation on the centroids of each segment. The resulting graph connects clusters that are close to each other. The intuition behind this is that neighboring clusters have a higher likelihood to belong to the same class, which is modeled with a statistical dependency in form of an edge in the CRF (see ► Fig. 54.17).

Assuming that the maximal clique size of the graph is 2, one can compute the conditional probability of the labels \mathbf{l} given the observations \mathbf{s} as:

$$p(\mathbf{l}|\mathbf{s}) = \frac{1}{Z(\mathbf{s})} \prod_{i=1}^N \varphi(\mathbf{s}_i, \mathbf{l}_i) \prod_{(i,j) \in \varepsilon} \psi(\mathbf{s}_i, \mathbf{s}_j, \mathbf{l}_i, \mathbf{l}_j), \quad (54.9)$$

where $Z(\mathbf{s}) = \sum_{\mathbf{l}} \prod_{i=1}^N \varphi(\mathbf{s}_i, \mathbf{l}_i) \prod_{(i,j) \in \varepsilon} \psi(\mathbf{s}_i, \mathbf{s}_j, \mathbf{l}_i, \mathbf{l}_j)$ is usually called the *partition function*, ε is the set of edges in the graph, and φ and ψ represent node and edge potentials. These potentials are positive functions that return large values if the labels \mathbf{l}_i correspond to the correct labels of the segments \mathbf{s}_i with a high probability. This means, that the potentials can be viewed as classifiers, where the node potential φ only uses the local information (i.e., features) of a segment for the classification, and the edge potential ψ measures the *consistency* of the labels between two adjacent segments. To define the node and edge potentials, simple rules can be used that relate the feature vector with the assigned class label, but it turns out that using a classifier such as AdaBoost (Freund and Schapire 1997)



■ Fig. 54.17

Object detection from 2D laser range scans using Conditional Random Fields (CRFs). The 2D data points shown here correspond to the scan depicted in ► Fig. 54.16. *Left:* First, the 2D scan points are clustered using jump-distance clustering. The *contoured points* correspond to the resulting clusters. *Right:* In a second step, a Delaunay triangulation is computed on the centroids of the clusters. This triangulation defines the structure of the CRF

for the node potentials φ and a simple rule for the edge potentials gives very good classification results (see Spinello et al. 2010). To achieve this, a *classification score* g_c is computed for each class based on the M weak classifiers h_i^c and their corresponding weight coefficients α_i^c as returned by the AdaBoost training algorithm:

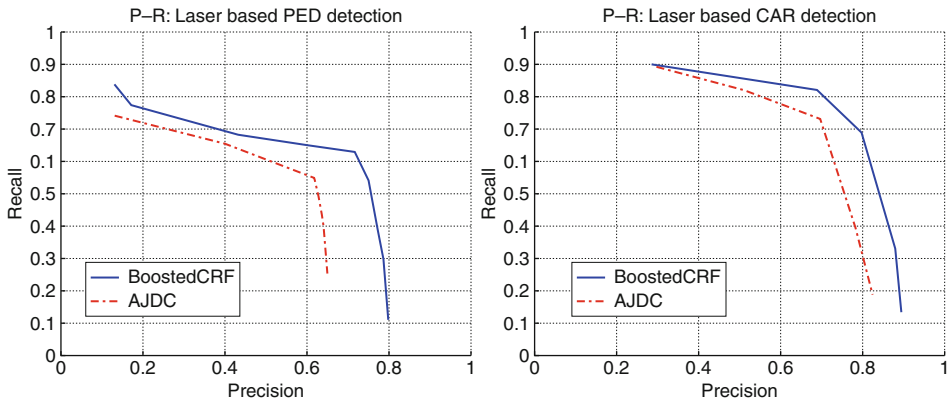
$$g_c(s_i) := \sum_{i=1}^M \alpha_i^c h_i^c(s_i). \quad (54.10)$$

To obtain a classification *likelihood*, the logistic function $\lambda(x) = (1 + e^{-x})^{-1}$ can be applied to g_c . For the edge potential, a good choice are rules such as “the closer two segments are to each other the higher is the chance that they correspond to the same class” or “the higher the classification score is for the two adjacent nodes on a given class, the higher is the probability that they have the same class labels.” For details about how this can be expressed in a mathematical formulation, see Spinello et al. (2010).

► Figure 54.18 shows precision-recall curves for results obtained with the boosted-CRF approach. As one can see, the additional information about the statistical dependence of labels from adjacent segments leveraged by the CRF approach effectively improves the classification results over the standard AdaBoost method.

4.1.3 Example of Detection by Classifier Ensemble

A feature-based fusion can be accomplished by a monolithic classifier method or an ensemble of classifiers. Particularly, in the latter case, the decisions are spread into small




■ Fig. 54.18

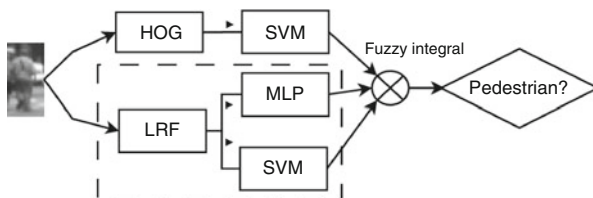
Classification results of pedestrians and cars from 2D laser range scans. The results obtained using a Conditional Random Field on node potentials using AdaBoost are compared with the results of an AdaBoost approach that is directly applied to the jump-distance clustered segments (AJDC). As can be seen, the boosted CRF performs better than AdaBoost alone (Results taken from Spinello et al. 2010)

specialized sub-modules, which are in charge of deciding over the feature representations. In this section, an ensemble fusion method, which performs over a set of features and image classifiers, is presented and discussed in a pedestrian detection task. In fact, concerning image classification, a single feature extractor-classifier is not usually able to deal with the diversity of multiple scenarios. Therefore, the integration of features and classifiers can bring benefits to cope with this problem, particularly when the parts are carefully chosen and synergistically combined.

Fusion of classifiers has been studied in the last few years with the aim of overcoming certain inabilities of single classifiers (Kuncheva 2004). The objective is to explore diversity of the component classifiers in order to enhance the overall classification performance. In other words, since there is no perfect individual classifier yet, by assembling them, one can complement the other. If there are errors in the ensemble components, it is expected that they occur on different image objects in order to give rise to fusion methods that improve the performance of the whole system. Additionally, the rationale of building ensembles is also that most individual classifiers agree in a certain way, such that the whole system can be more successful than its component parts.

From experimental studies, a set of feature extractors and strong classifiers were chosen which could interact together to improve the general performance of the final classification system. In this sense, the choice of the feature extractors, histogram of oriented gradients (HOG) and local receptive fields (LRF), was motivated by the studies found in Dalal and Triggs (2005), Munder and Gavrila (2006), and Szarvas et al. (2006). Dalal and Triggs (2005) presented an experimental analysis demonstrating that HOG features outperform PCA-SIFT, Haar wavelets, and shape contexts in a complex dataset. Munder and Gavrila (2006) also experimentally showed that LRF features present superior performance in comparison with PCA and Haar wavelets, although computed from an MLP over Haar wavelets, or PCA features, and classified by SVM or Adaboost, which turned it to a method more sensitive to lighting variations. Szarvas et al. (2006) found that LRFs built on CNNs have great potential in pedestrian recognition. Our goal is thus to show that there is an opportunity to integrate synergistically the outputs of high-performance classifiers performing over these two types of features.

After an extensive experimental work, the final architecture of the proposed synergistic method was built, and is illustrated in  Fig. 54.19 [see Oliveira et al. (2010a) for the complete description of the ensemble evaluation]. In the feature level, HOG and LRF feature extractors are in charge of representing the objects in different ways. It was demonstrated that HOG are better to represent pedestrians, while LRF are better to represent background. Actually, they complement each other, in many circumstances with respect to pedestrian/non-pedestrian classification ability. Instead of employing weak classifiers in the fashion of boosting methods, strong classifiers were used in order to explore the complementarity of the features. It means that once the chosen features provide synergism in the way of acting, the lesser errors they commit individually, the better the integration made by the fusion method. The name of the method was coined as HLSM-FINT, standing for each initial letter of the feature extractors and classifiers used in the ensemble.



■ Fig. 54.19

Classifier ensemble synergism: The component classifiers, SVM and MLP, run over the feature extractors HOG and LRF. After a thorough evaluation of a bunch of methods, a Sugeno fuzzy integral method was chosen as the best one to combine the classifier results

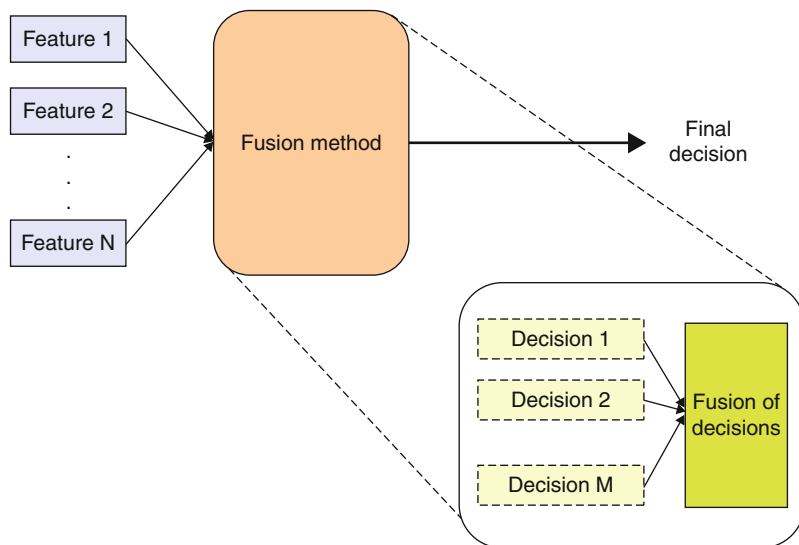
HLSM-FINT was evaluated, by means of receiver operating characteristic (ROC) curves, over DaimlerChrysler Munder and Gavrila (2006), INRIA Dalal and Triggs (2005), and Nature Inspired Smart Information System (NISIS) competition datasets Oliveira et al. (2010a). Furthermore, the DaimlerChrysler datasets were changed to incorporate some usual situations involving lighting and contrast changes. At the end, HLSM-FINT presented an averaged performance of 94% of hit rate (HR) with 3% of false alarm rate (FAR) over all datasets evaluated.

4.2 Detection by Sensor Fusion

Actually, in the last decades, several researchers have been developing complete perception architectures for intelligent transportation systems (ITS) (Reisman et al. 2004; Leibe et al. 2005, 2007).

In real life, perception systems have become a reality with the deployment of camera-based object detection and lane departure systems, in top-of-the-line vehicles. Although these systems are intended to aid the driver in hazardous situations, much has yet to be done in order to make them completely reliable in several circumstances. In this regard, multi-sensor architectures may bring complementarity and redundancy, making perception systems more robust. The idea is to build a system to take advantage of the strengths of various sensors.

For intelligent vehicle applications, some sensors appear as the best choices to perceive the environment in order to provide sensing information and intelligent decisions. They are range finders like lasers and cameras. With those sensors, there are many approaches with the aim of getting the best of sensor data integration. For instance, Perrollaz et al. (2010a) propose a probabilistic representation of the uncertainty of the fusion of two cameras (stereovision); they build this representation into a grid framework for object detection. Oliveira et al. (2010) suggest that a semantic fusion of laser and vision in a feature level is highly advantageous for object detection, if it is possible to decouple the detection dependency of the vision system with respect to the laser detection Oliveira and Nunes (2010). Scheunert et al. (2008) integrate a stereovision with a laser scanner for



■ Fig. 54.20

Feature-based framework. Features 1, ..., N are extracted in such a way that they could represent the object of interest in different manners. After that, the features are fused in the fusion method, which is ultimately structured as a monolithic decision module or a set of specialized decision sub-modules integrated in a more generic method

obstacle avoidance; that work is also structured in an occupancy grid framework as in Perrollaz et al. (2010b). Spinello et al. (2009b) propose a cooperative fusion of independent sensor detection of a laser scanner and a vision system; both sensors are fused by means of a tracking system used in each sensor space with respective data association.

A feature-based fusion is characterized by an integration in the feature level. One should expect a general feature-based fusion framework depicted in Fig. 54.20 (for a broader view of various fusion architectures, see Dasarathy 1997).

In Fig. 54.20, it is shown how the feature-based fusion is accomplished. The figure depicts a general framework regardless of where the features are coming. Indeed, features may be obtained from a unique sensor or multiple ones. Also, the features can represent an object in different sensor spaces (for instance, laser and camera). For all those reasons, the choice of the fusion method is of underlying importance, and the fusion architecture ought to be designed taking into account the different characteristics brought for each feature extractor.

In turn, the feature extractors are in charge of getting good representations of an object of interest in the scene. For “good,” one may expect that a particular object should be represented differently from other objects pertaining to a different category. Unfortunately, obtaining this unique representation for each different type of an object is cumbersome, if not impossible, which means that errors will be encountered. Those errors will be dealt with in the fusion method, which will later decide, in a certain level

(monolithically or via a classifier ensemble methods), how to choose the best representation among the inputs (features $1, \dots, N$).

The choice of the fusion methods can be made namely by considering the performance of the specialized decision sub-modules (if they are presented in the architecture), the characteristics of the feature extractors, and/or how much information each feature carries on.

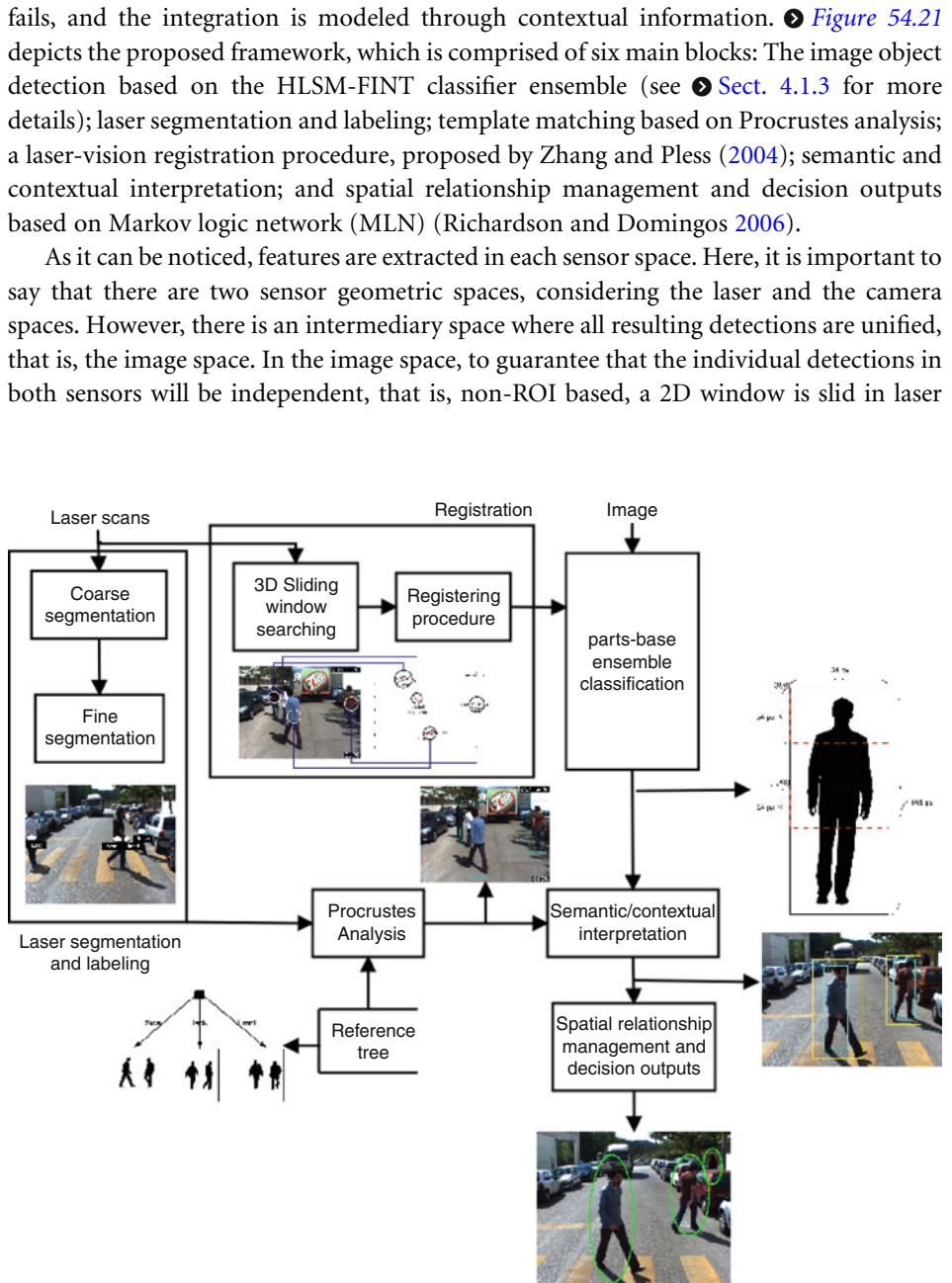
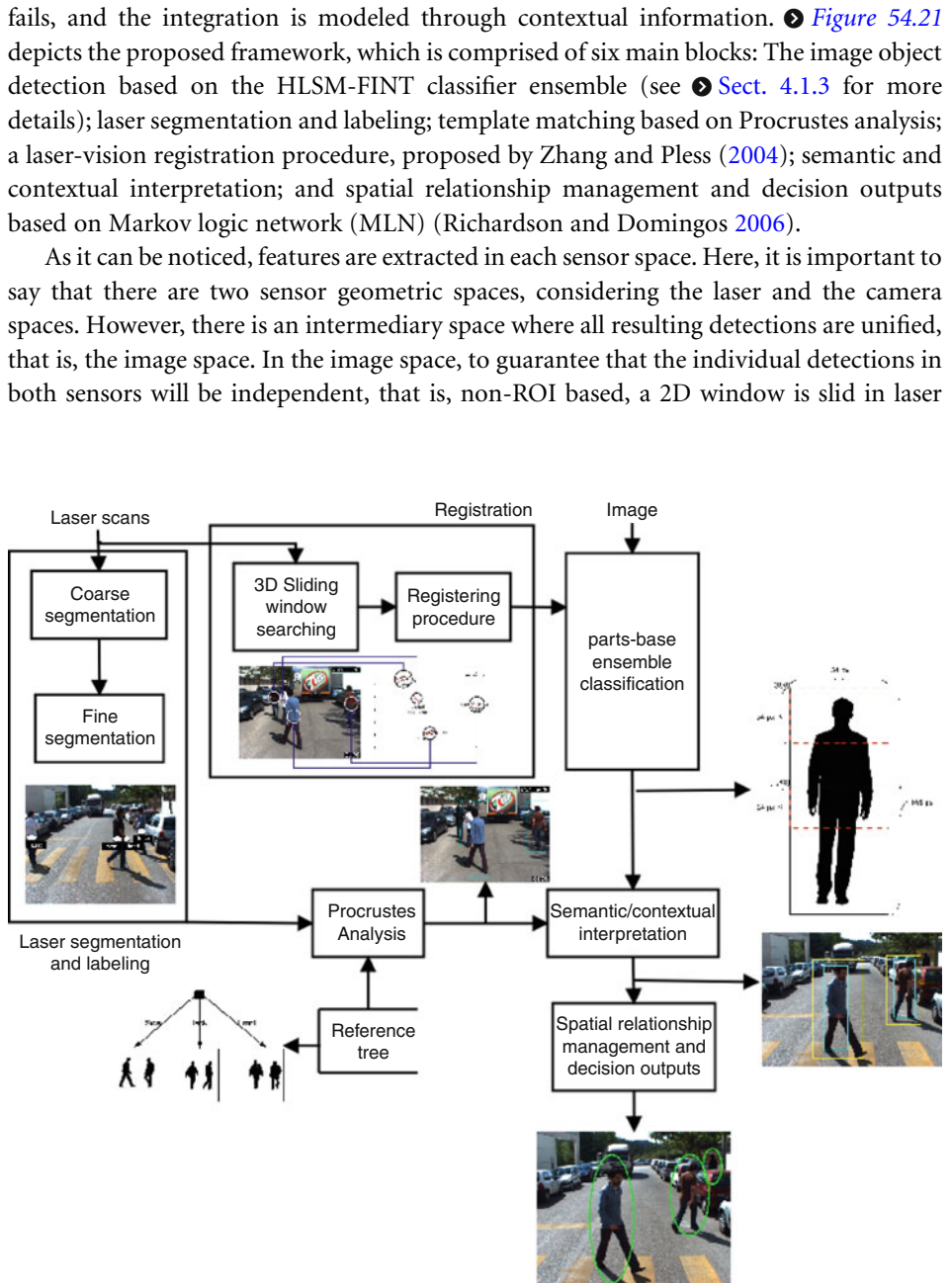
Another aspect of a fusion method design is the information that each feature or the specialized fusion sub-modules brings with them. The way to deal with this feature inherent information is crucial since it is expected that a fusion method explores not only redundancy but also complementarity of information; in the first case, the fusion will try to match similar characteristics that reinforces the decision, while in the latter one, the fusion method appeals to the fact that one feature representation can complement information missing in the others. The goal for the complementarity is also to decrease the decision uncertainty by building the final decision from small pieces of information coming from multiple sources or representations. Next, a sensor fusion method based on semantic information of parts-based detectors is described.

Semantic fusion: So far, fusion of laser scanner and vision sensors has been performed by assuming that the probability to find an object is identical and usually independent, in both sensor spaces. There are two main approaches for this type of fusion:

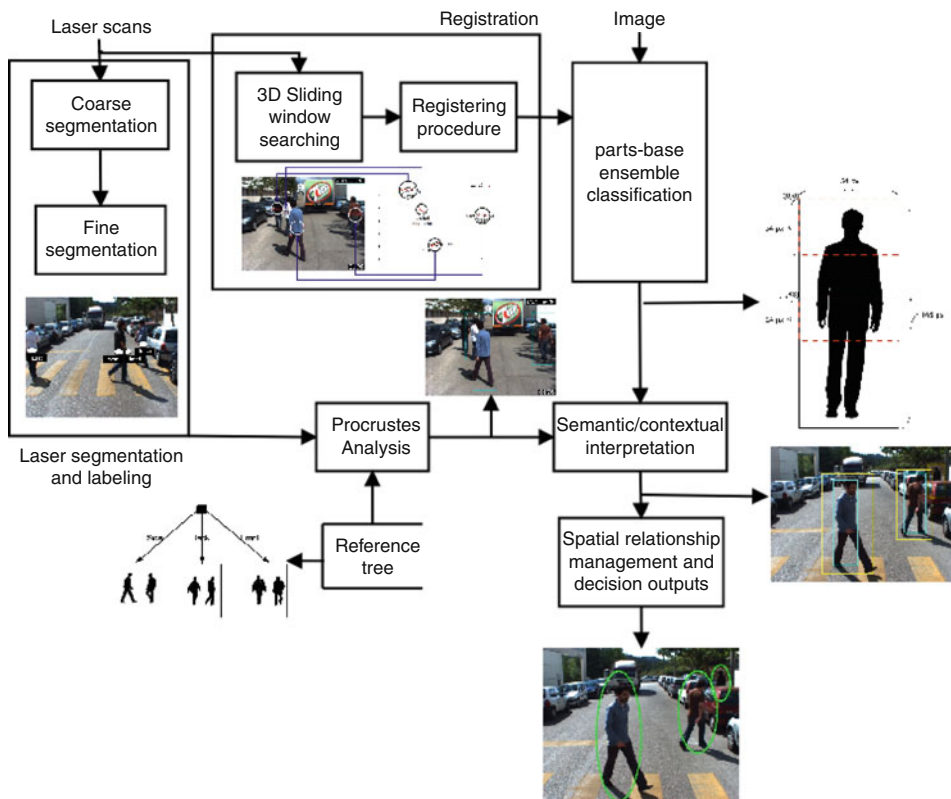
- A laser scanner segmentation method is used to find most likely regions of interest (ROIs), where an image classification system is applied.
- Identically and independent distribution(IID) integration of sensor-driven classifiers (posterior probabilities), or sensor-driven features.

Some of the works based on these ROI-driven methods are presented as follows. Szarvas et al. (2006) rely entirely on laser ROIs in order to find probable areas where a pedestrian might be. Each image projected ROI is classified by a convolutional neural network (CNN) Lecun et al. (1998). When the laser fails, no pedestrians are detected. Broggi et al. (2008) propose an on-spot pedestrian classification system, which first uses a laser scanner to determine areas between vehicles, and then a Haar-like feature/Adaboost classification system to name the projection of these areas in the image. Mahlich et al. (2006) propose a spatio-temporal alignment to integrate the laser scanner and the monocular camera. For that purpose, features in both sensor spaces are extracted, feeding a Bayesian classifier to detect cars. Douillard et al. (2007) propose an approach similar to the previous one, but rather than a Bayesian rule, they use a conditional random field conditional random fields (CRF). Additionally, they are able not only to classify laser and image features, but also to find a temporal relationship between features of sequential sensor readings. Premebida et al. (2007) propose to classify features in laser space with a Gaussian mixture model (GMM) while Haar-like features are extracted from the image objects and classified by an Adaboost. The confidence scores of the classifiers feed a Bayesian rule to make the final decision. The goal is to classify vehicles and pedestrians.

Unlike the aforementioned approaches, the semantic fusion method Oliveira et al. (2010b) deals with partial segments, it is able to recover depth information even if the laser

fails, and the integration is modeled through contextual information.  Figure 54.21 depicts the proposed framework, which is comprised of six main blocks: The image object detection based on the HLSM-FINT classifier ensemble (see  Sect. 4.1.3 for more details); laser segmentation and labeling; template matching based on Procrustes analysis; a laser-vision registration procedure, proposed by Zhang and Pless (2004); semantic and contextual interpretation; and spatial relationship management and decision outputs based on Markov logic network (MLN) (Richardson and Domingos 2006).

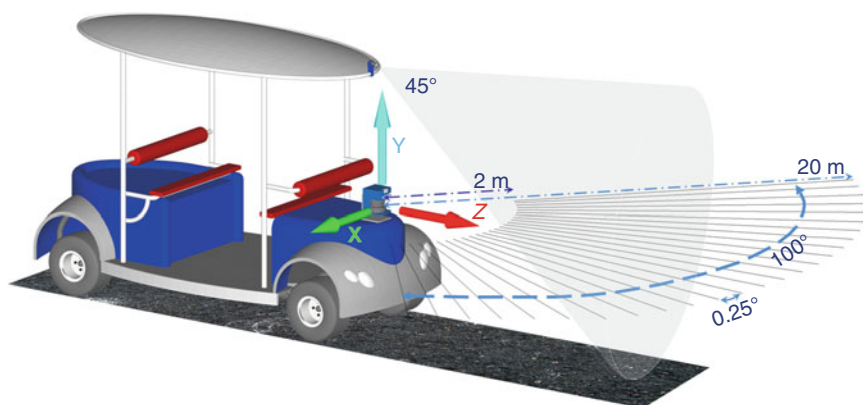
As it can be noticed, features are extracted in each sensor space. Here, it is important to say that there are two sensor geometric spaces, considering the laser and the camera spaces. However, there is an intermediary space where all resulting detections are unified, that is, the image space. In the image space, to guarantee that the individual detections in both sensors will be independent, that is, non-ROI based, a 2D window is slid in laser



■ Fig. 54.21

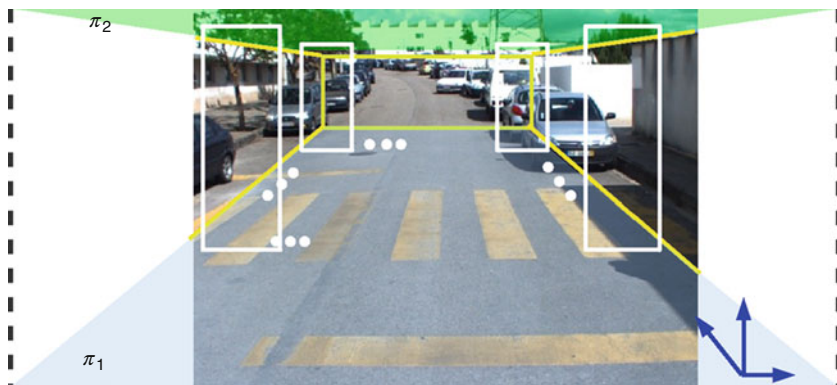
The proposed framework is composed of six main blocks: The image object detection based on a parts-based ensemble detector using HLSM-FINT; laser segmentation and labeling; template matching; laser-vision registration, based on Zhang and Pless' method Zhang and Pless (2004); semantic and contextual interpretation; and the semantic fusion based on MLN

space, and subsequently projected into the image space. This has threefold advantages: the economy on windows being slid in various scales and positions in the image space; a 3D control of the vision detector, which allows a depth information even if the laser fails; and, finally, the projections of the laser segments can also be treated independent in the same projections, in the image space. For those sliding windows in the 3D laser space, it is assumed that the laser is always parallel to the ground (see ▶ Fig. 54.22). ▶ Figure 54.23 illustrates the geometry of this process.



■ Fig. 54.22

Setup vehicle: The perception system ranges from 2 up to 20 m in both sensor spaces. A SICK 2D laser scanner (100° of aperture angle) and a Point Grey camera set to a resolution of $1,024 \times 768$ pixels (45° of field of view) were used

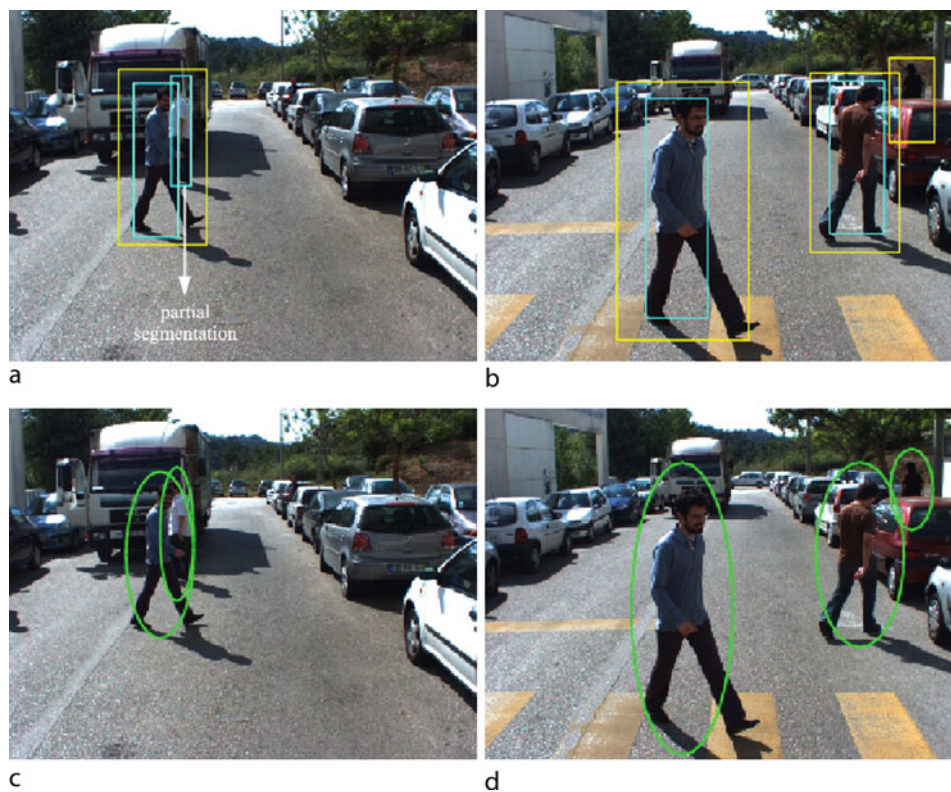


■ Fig. 54.23

A window sized $1.0 \text{ m} \times 1.8 \text{ m}$ is shifted onto horizontal and vertical directions in laser sensing space, with a stride of 0.20 m , ranging over 2 m up to 20 m in depth. The searching area is constrained by the viewport formed by the planes π_1 and π_2

The segmentation, in laser space, uses a featureless approach. The rationale of this method is to segment coarsely the laser raw points. These coarse segments, $\{c_n\}$, where $n = 1, \dots, N$, are posteriorly sub-segmented into fine segments, $\{f_m\}$, where $m = 1 \dots 3$. This latter step is done by a β -skeleton random graph. As the main goal is to detect pedestrians in outdoor, the laser is mounted at the pedestrian-waist level (see [Fig. 54.22](#)), avoiding common problems as those observed in leg-level mounted systems. Therefore, the fine segmentation step expects at most three body parts (two arms and the torso).

On the other hand, in image space, the HLSM-FINT is applied in a parts-based detector. Actually, the INRIA datasets Dalal and Triggs (2005) were used to train a two-part detector, which classifies the hip region and the shoulder-head region of the human body. The idea is to allow the image detector to cope with partial occlusion in the same way





■ Fig. 54.24

Samples of the semantic fusion in action. In (a) and (b), results of the individual classification in each sensor space (the *outer rectangular bounding box* is given by an image detector, while the inner rectangular box (or two boxes in a) is given by a laser segmentation). In (b), an example where the laser fails and the image detector complements the recognition process. Note that the depth information when laser fails is estimated from the 3D searching windows. In (c), the result of the semantic fusion (*left to right*) from (a) and (b), respectively

that the laser detector does. With that, there are parts being labeled in image and laser spaces. Each of these parts is then contextually interpreted, and finally modeled by an MLN.

In MLN framework (Richardson and Domingos 2006), the information of the parts is modeled according to first-order clauses. The first-order rules are then structured as a Markov random fields (MRF) with a grounded first-order formula in each node, which is so called a ground MRF. After the inference process over the MRF, marginal probabilities are then computed and out-putted for each test case. To build the MLN, the Alchemy library was used, available at <http://alchemy.cs.washington.edu/>.

As mentioned before, the semantic fusion of laser and vision differs from previous fusion approaches on these sensors, in the following twofold reasons: non-ROI-driven and non-IID based fusion. These characteristics led to a very robust system with independent and synergistic sensor fusion, based on a semantic approach (see  Fig. 54.21).

The rationale of the semantic fusion system is based on parts-based classifiers in both sensor spaces, which are integrated with the detachment of the vision detector from the laser detector. This independent fusion was achieved by the 2D sliding window running in 3D laser space and the processing of the object parts in a non-IID framework, that is, MLN. With that, it was possible to make the vision detection system to run regardless of the laser hypothesized ROI. The validation of the semantic method was made over collected datasets, which are available at <http://www.isr.uc.pt/~lreboucas>. Some results of the method application over the collected image frames are depicted in  Fig. 54.24.

5 Conclusions

This chapter provided an overview of the DATMO problem for autonomous driving and outlined three main classes of solutions: traditional, model based, and grid based.

The traditional approach consists of data segmentation, association, and filtering steps. Recent innovations within this class are led by pattern recognition techniques, which are capable of recognizing scene participants (e.g., pedestrians and vehicles) from a single frame of data. These techniques are particularly useful for identifying potentially dynamic objects when these objects are not yet moving.

The model-based approach is able to infer and use additional shape knowledge, which is particularly useful when working with large objects such as cars, buses, and trucks. These objects are often split up into separate clusters due to occlusion, making segmentation and association of data difficult. Using geometric models, the model-based approach bypasses segmentation and association steps altogether. It also leads to more accurate motion estimation, which is particularly important for objects with non-holonomic motion and/or objects moving at high speeds.

The grid-based approach delays data association decisions to later stages of the DATMO pipeline. At early stages, it constructs a low level grid representation capturing occupancy and velocity estimates of grid cells. This representation can be used directly for motion planning in dynamic free space or supplemented by segmentation and association techniques to identify individual targets. The grid-based approach is particularly useful

for identifying a broad variety of objects regardless of their appearance. This makes it capable of identifying small participants such as young children and pets.

DATMO for autonomous driving is a young and rapidly developing field. Although many interesting and successful methods have been developed, a fully reliable autonomous DATMO system is yet to be built. For now, existing DATMO systems cannot rival human capabilities in terms of detection range. Existing DATMO approaches have not yet been shown to be fully reliable with respect to participants of unusual appearance, for example, unusual vehicles (e.g., construction equipment or parade floats), people wearing extravagant costumes or fully draped in loose clothing, animals, or people pushing various objects. Performance of DATMO techniques in adverse weather or lighting conditions is yet to be studied in depth.

References

- Andrieu C, De Freitas N, Doucet A, Jordan M (2003) An introduction to MCMC for machine learning. *Mach Learn* 50(1):5–43
- Arras K, Mozos O, Burgard W (2007) Using boosted features for the detection of people in 2d range data. In: *Proceedings of IEEE international conference on robotics and automation (ICRA)*, Rome, pp 3402–3407
- Arulampalam S, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. *IEEE Trans Signal Process* 50:174–188
- Badino H, Franke U, Mester R (2007) Free space computation using stochastic occupancy grids and dynamic programming. In: *IEEE international conference on computer vision, workshop on dynamical vision*, Rio de Janeiro
- Bar-Shalom Y, Jaffer A (1972) Adaptive nonlinear filtering for tracking with measurements of uncertain origin. In: *Conference on decision and control and 11th symposium on adaptive processes*, New Orleans. Institute of Electrical and Electronics Engineers, New York, pp 243–247
- Bertozzi M, Broggi A, Castelluccio S (1997) A real-time oriented system for vehicle detection. *J Syst Archit* 43:317–325
- Bertozzi M, Broggi A, Fascioli A (2000) Vision-based intelligent vehicles: state of the art and perspectives. *Robot Auton Syst* 32:1–16
- Blackman S, Co R, El Segundo C (2004) Multiple hypothesis tracking for multiple target tracking. *IEEE Aerosp Electron Syst Mag* 19(1 Part 2):5–18
- Blackwell D (1947) Conditional expectation and unbiased sequential estimation. *Ann Math Stat* 18(1):105–110
- Borgefors G (1986) Distance transformations in digital images. *Comput Vision Graph Image Process* 34:344–371
- Braillon C, Pradalier C, Usher K, Crowley J, Laugier C (2006) Occupancy grids from stereo and optical flow data. In: *Proceedings of international symposium on experimental robotics*, Rio de Janeiro
- Broggi A, Cerri P, Ghidoni S, Grisleri P, Gi J (2008) Localization and analysis of critical areas in urban scenarios. In: *Proceedings of IEEE international symposium on intelligent vehicles*, Eindhoven, pp 1074–1079
- Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. *IEEE Trans Pattern Anal Mach Intell* 24:603–619
- Coué C, Pradalier C, Laugier C, Fraichard T, Bessiére P (2006) Bayesian occupancy filtering for multitarget tracking: an automotive application. *Int J Robot Res* 25(1):19–30
- Cox I (1993) A review of statistical data association techniques for motion correspondence. *Int J Comput Vis* 10(1):53–66
- Cucchiara R, Piccardi M (1999) Vehicle detection under day and night illumination. In: *Proceedings of the 3rd international ICSC symposium on intelligent industrial automation*, Genova
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, vol 1, San Diego, pp 886–893

- Dasarathy B (1997) Sensor fusion potential exploitation – innovative and illustrative applications. In: Proceedings of the IEEE special issue on sensor fusion, vol 85, pp 24–38. IEEE
- Diard J, Bessiere P, Mazer E (2003) A survey of probabilistic models using the bayesian programming methodology as a unifying framework. In: The second international conference on computational intelligence, robotics and autonomous systems (CIRAS 2003), Singapore
- Dickmanns E, Behringer R, Dickmanns D, Hildebrandt T, Maurer M, Thomaneck F, Schiehlen J (1994) The seeing passenger car “vamos-p”. In: Proceedings of the intelligent vehicles 1994 symposium, Paris, France, pp 68–73
- Dollar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Miami, USA, pp 304–311
- Douillard B, Fox D, Ramos F (2007) A spatio-temporal probabilistic model for multi-sensor object recognition. In: Proceedings of IEEE/RSJ international conference on intelligent robots and systems, San Diego, CA, USA, pp 2402–2408
- Douillard B, Fox D, Ramos F (2008) Laser and vision based outdoor object mapping. In: Proceedings of robotics: science and systems (RSS), Zurich
- Enzweiler M, Gavrila D (2009) Monocular pedestrian detection: survey and experiments. IEEE Trans Pattern Anal Mach Intell 31(12):2179–2195
- Ess A, Leibe B, Schindler K, Van Gool L (2009) Moving obstacle detection in highly dynamic scenes. In: Proceedings of IEEE international conference on Robotics and Automation (ICRA), Kobe, Japan
- Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Anchorage, Alaska, USA, pp 1–8
- Fod A, Howard A, Mataric M (2002) Laser-based people tracking. In: Proceedings of IEEE international conference on robotics and automation (ICRA), vol 3, Washington, DC, pp 3024–3029
- Fortmann T, Bar-Shalom Y, Scheffe M (1983) Sonar tracking of multiple targets using joint probabilistic data association. IEEE J Ocean Eng 8(3):173–184
- Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
- Gavrila D (2000) Pedestrian detection from a moving vehicle. In: European conference on computer vision (ECCV), Dublin, pp 37–49. IEEE Computer Society
- Gordon N (1993) Bayesian methods for tracking. PhD thesis, University of London
- Hastings WK (1970) Monte Carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109
- Kleinhagenbrock M, Lang S, Fritsch J, Lömker F, Fink G, Sagerer G (2002) Person tracking with a mobile robot based on multi-modal anchoring. In: IEEE international workshop on robot and human interactive communication (ROMAN), Berlin, Germany
- Kuehnle A (1991) Symmetry-based recognition of vehicle rears. Pattern Recognit Lett 12:249–258
- Kuncheva L (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Hoboken
- Labayrade R, Aubert D, Tarel J (2002) Real time obstacles detection on non flat road geometry through v-disparity representation. In: Proceedings of IEEE Intelligent Vehicle Symposium (IV), Versailles, France
- Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmentation and labeling sequence data. In: International conference on machine learning (ICML), Williamstown, pp 282–289
- Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2279–2324
- Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), vol 1, San Diego, pp 878–885
- Leibe B, Cornelis N, Cornelis K, Van Gool L (2007) Dynamic 3D scene analysis from a moving vehicle. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), vol 1, Minneapolis, pp 1–8
- Leonard J, How J, Teller S, Berger M, Campbell S, Fiore G, Fletcher L, Frazzoli E, Huang A, Kara-man S et al (2008) A perception-driven autonomous urban vehicle. J Field Robot 25(10):727–774
- Lundquist C, Schon T (2008) Road geometry estimation and vehicle tracking using a single track model. In: Intelligent vehicles symposium, 2008

- IEEE, pp 144–149. IEEE, Eindhoven, The Netherlands
- MacKay DJC (1998) Introduction to Monte Carlo methods. In: Jordan MI (ed) *Learning in graphical models*, NATO science series. Kluwer Academic, Dordrecht, pp 175–204
- Mahlisch RS, Ritter W, Dietmayer K (2006) Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection. In: *Proceedings of IEEE international symposium on intelligent vehicles*, Tokyo, Japan, pp 424–429
- Matthies L, Elfes A (1988) Integration of sonar and stereo range data using a grid-based representation. In: *Proceedings of IEEE international conference on robotics and automation*, Philadelphia
- Mekhnacha K, Mao Y, Raulo D, Laugier C (2008) Bayesian occupancy filter based Fast Clustering-Tracking algorithm. In: *Proceedings of IEEE/RSJ international conference on intelligent robot and systems*, Nice
- Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E et al (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21(6):1087
- Mikolajczyk K, Schmid C, Zisserman A (2004) Human detection based on a probabilistic assembly of robust part detectors. In: *European conference on computer vision (ECCV)*, Prague, Czech Republic, pp 69–82
- Mohan A, Papageorgiou C, Poggio T (2001) Example-based object detection in images by components. *IEEE Trans Pattern Anal Mach Intell* 23(4):349–361
- Montemerlo M (2003) FastSLAM: a factored solution to the simultaneous localization and mapping problem with unknown data association. PhD thesis, Robotics Institute, Carnegie Mellon University
- Montemerlo M, Becker J, Bhat S, Dahlkamp H, Dolgov D, Ettinger S, Hähnel D, Hilden T, Hoffmann G, Huhnke B, Johnston D, Klumpp S, Langer D, Levandowski A, Levinson J, Marcil J, Orenstein D, Paefgen J, Penny I, Petrovskaya A, Plüger M, Stanek G, Stavens D, Vogt A, Thrun S (2008) Junior: the stanford entry in the urban challenge. *J Field Robot* 25(9):569–597
- Moravec H (1988) Sensor fusion in certainty grids for mobile robots. *AI Mag* 9(2)
- Munder S, Gavrilu D (2006) An experimental study on pedestrian classification. *IEEE Trans Pattern Anal Mach Intell* 28:1863–1868
- Murphy K, Russell S (2001) *Rao-blackwellized particle filtering for dynamic Bayesian networks*. Springer, Heidelberg
- Murray D, Little J (2000) Using real-time stereo vision for mobile robot navigation. *Auton Robot* 8(2):161–171
- Oliveira L, Nunes U (2010) Context-aware pedestrian detection using lidar. In: *Proceedings of IEEE international symposium on intelligent vehicles*, San Diego, CA, USA
- Oliveira L, Nunes U, Peixoto P (2010a) On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Trans Intell Transp Syst* 11:16–27
- Oliveira L, Nunes U, Peixoto P, Silva M, Moita F (2010b) Semantic fusion of laser and vision in pedestrian detection. *Pattern Recognit* 43:3648–3659
- Papageorgiou C, Poggio T (2000) A trainable system for object detection. *Int J Comput Vis* 38(1):15–33
- Perrollaz M, Spalanzani A, Aubert D (2010a) Probabilistic representation of the uncertainty of stereo-vision and application to obstacle detection. In: *Proceedings of IEEE international symposium on intelligent vehicles*, San Diego, CA, USA, pp 313–318
- Perrollaz M, Yoder J-D, Laugier C (2010b) Using obstacle and road pixels in the disparity space computation of stereovision based occupancy grids. In: *Proceedings of IEEE international conference on intelligent transportation systems*, Madeira, Portugal
- Petrovskaya A (2011) Towards dependable robotic perception. Ph D thesis, Stanford University, Stanford
- Petrovskaya A, Khatib O (2011) Global localization of objects via touch. *IEEE Trans Robot* 27(3):569–585
- Petrovskaya A, Thrun S (2009) Model based vehicle detection and tracking for autonomous urban driving. *Auton Robot* 26(2):123–139
- Premebida C, Monteiro G, Nunes U, Peixoto P (2007) A lidar and vision-based approach for pedestrian and vehicle detection and tracking. In: *Proceedings of IEEE international conference on intelligent transportation systems*, Seattle, USA, pp 1044–1049
- Prisacariu VA, Reid I (2009) fasthog- a realtime gpu implementation of hog technical report no. 2310/09
- Rao C (1945) Information and accuracy obtainable in one estimation of a statistical parameter. *Bull Calcutta Math Soc* 37:81–91

- Reid DB (1979) An algorithm for tracking multiple targets. *IEEE Trans Autom Control* 24:843–854
- Reisman P, Mano O, Avidan S, Shashua A (2004) Crowd detection in video sequences. In: *Proceedings of IEEE international symposium on intelligent vehicles*, Parma, Italy, pp 66–71. IEEE
- Richardson M, Domingos P (2006) Markov logic networks. *Mach Learn* 62:107–136
- Richter E, Schubert R, Wanielik G (2008) Radar and vision based data fusion-advanced filtering techniques for a multi object vehicle tracking system. In: *Proceedings of intelligent vehicles symposium, 2008 IEEE*, Eindhoven, The Netherlands pp 120–125. IEEE
- Scheunert U, Mattern N, Lindner P, Wanielik G (2008) Generalized grid framework for multi sensor data fusion. *J Inf Fusion*, 814–820
- Schulz D, Burgard W, Fox D, Cremers AB (2003) People tracking with mobile robots using sample-based joint probabilistic data association filters. *Int J Robot Res (IJRR)* 22(2):99–116
- Sittler R (1964) An optimal data association problem in surveillance theory. *IEEE Trans Militar Electron* 8(2):125–139
- Spinello L, Macho A, Triebel R, Siegwart R (2009a) Detecting pedestrians at very small scales. In: *Proceedings of IEEE international conference on intelligent robots and systems (IROS)*, St. Louis, pp 4313–4318
- Spinello L, Triebel R, Siegwart R (2009b) A trained system for multimodal perception in urban environments. In *proceedings of the workshop on people detection and tracking of IEEE ICRA 2009*, Kobe, Japan
- Spinello L, Triebel R, Siegwart R (2010) Multi-class multimodal detection and tracking in urban environments. *Int J Robot Res (IJRR)* 29(12):1498–1515
- Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: a review. *IEEE Trans Pattern Anal Mach Intell* 28:694–711
- Szarvas M, Sakai U, Ogata J (2006) Real-time pedestrian detection using lidar and convolutional neural networks. In: *Proceedings of IEEE international symposium on intelligent vehicles*, Tokyo, Japan, pp 213–218
- Thrun S, Burgard W, Fox D (2005) *Probabilistic robotics*. MIT Press, Cambridge, MA
- Tuzel O, Porikli F, Meer P (2007) Human detection via classification on riemannian manifolds. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, Rio de Janeiro, Brazil, pp 1–8
- Vermaak J, Godsill S, Perez P (2005) Monte Carlo filtering for multi target tracking and data association. *IEEE Trans Aerosp Electron Syst* 41(1):309–332
- Viola P, Jones MJ, Snow D (2003) Detecting pedestrians using patterns of motion and appearance. In: *Proceedings of IEEE international conference on computer vision (ICCV)*, vol 2, Nice, pp 734–741
- Vu T (2009) *Vehicle perception: localization, mapping with detection, classification and tracking of moving objects*. PhD thesis, Institut National Polytechnique De Grenoble
- Vu T, Aycard O (2009) Laser-based detection and tracking moving objects using data-driven markov chain Monte Carlo. In: *Proceedings of IEEE international conference on robotics and automation (ICRA 2009)*, Kobe, Japan, pp 3800–3806. IEEE
- Wang X, Han TX, Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: *Proceedings of 2009 IEEE 12th international conference on computer vision*, Kyoto, Japan, pp 32–39
- Wu B, Nevatia R (2005) Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: *Proceedings of IEEE international conference on computer vision (ICCV)*, Beijing, China
- Xavier J, Pacheco M, Castro D, Ruano A, Nunes U (2005) Fast line, arc/circle and leg detection from laser scan data in a player driver. In: *Proceedings of IEEE international conference on robotics and automation (ICRA)*, Barcelona, pp 3930–3935
- Zhang Q, Pless R (2004) Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: *Proceedings of IEEE/RSJ international conference on intelligent robots and systems*, Sendai, Japan, vol 3, pp 2301–2306
- Zhu Q, Yeh MC, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, vol 2, New York, pp 1491–1498

55 Iterative Motion Planning and Safety Issue

Thierry Fraichard¹ · Thomas M. Howard²

¹INRIA Grenoble - Rhône-Alpes, CNRS-LIG and Grenoble University, Grenoble, France

²Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

1	Introduction	1435
1.1	Hierarchical Navigation	1436
1.2	Constraints	1436
1.3	Nomenclature	1437
2	Motion Safety	1437
2.1	Inevitable Collision States	1437
2.2	Motion Safety Criteria	1439
2.2.1	Limited Decision Time	1439
2.2.2	Reasoning About the Future	1440
2.2.3	Appropriate Lookahead	1440
2.3	Future Modeling	1440
2.3.1	Deterministic Models	1441
2.3.2	Conservative Models	1441
2.3.3	Probabilistic Models	1441
2.4	Literature Review	1442
3	Iterative Motion Planning	1443
3.1	Potential Field Techniques	1443
3.2	Sampling Techniques	1443
3.2.1	Input Space Sampling	1444
3.2.2	State Space Sampling	1446
3.3	Graph Search Techniques	1449
3.3.1	Ego-graphs	1449
3.3.2	Rapidly Exploring Dense Trees	1450
3.3.3	Recombinant State Lattices	1451

4 *Applications* 1451

4.1 Roadway and Urban Navigation 1451

4.2 Off-Road Navigation 1453

5 *Conclusions and Further Reading* 1454

Abstract: This chapter addresses safe mobile robot navigation in complex environments. The challenges in this class of navigation problems include nontrivial vehicle dynamics and terrain interaction, static and dynamic environments, and incomplete information.

This complexity prompted the design of hierarchical solutions featuring a multilevel strategy where strategic behaviors are planned at a global scale and tactical or safety decisions are made at a local scale. While the task of the high level is generally to compute the sequence of waypoints or waystates to reach the goal, the local planner computes the actual trajectory that will be executed by the system. Due to computational resource limitations, finite sensing horizon, and temporal constraints of mobile robots, the local trajectory is only partially computed to provide a motion that makes progress toward the goal state. This chapter focuses on safely and efficiently computing the local trajectory in the context of mobile robot navigation.

This chapter is divided into three sections: motion safety, iterative motion planning, and applications. Motion safety discusses the issues related to determining if a trajectory is safely traversable by a mobile robot. Iterative motion planning reviews developments in local motion planning search space design with a focus on potential field, sampling, and graph search techniques. The applications section surveys experiments and applications in autonomous mobile robot navigation in outdoor and urban environments.

1 Introduction

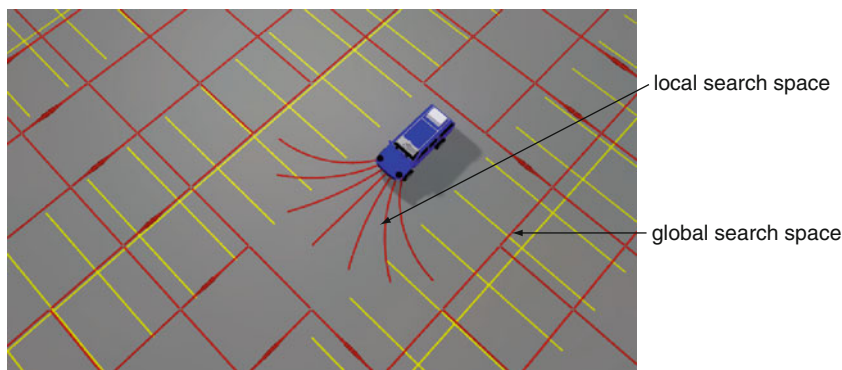
The four most significant challenges in autonomous mobile robot navigation are safety, feasibility, optimality, and efficiency. Safety involves perceiving and understanding the complicated interaction between the vehicle and its surroundings to identify hazardous and harmless regions. Feasibility, optimality, and efficiency are issues related to planning motions through the observed environment. Feasibility represents the quality of the mapping between vehicle inputs (actions) and predicted state space response (trajectories). Optimality is a metric measuring the quality of the planned trajectory. This is often measured with respect to the global optimal solution, which in realistic environments can only be nearly approximated with perfect prior information at enormous computational expense. Efficiency is the computational expense of calculating a trajectory. This is particularly important for mobile robot applications because computational resources are limited and iterative motion planners must react in real time to vehicle and environmental changes. An ideal iterative motion planner would always determine a feasible, optimal trajectory through the environment while consuming nearly no computational resources or time. This chapter discusses applied research in motion safety and iterative motion planning with application to mobile robot navigation in constrained, cluttered, and unstructured environments.

1.1 Hierarchical Navigation

Hierarchical mobile robot navigation approaches are appropriate when the goal state is beyond the sensor horizon, when the environment is dynamic, or when computational resources are limited. Planning a high-fidelity trajectory through an incomplete or unknown environment may not lead to more intelligent navigation decisions because it is no more informed than simpler, more efficient approximation of vehicle motion. A hierarchical navigation architecture combines expressive, high-fidelity local search with coarse, approximate global motion planning (► Fig. 55.1). In this example, the local search (represented by a series of arcs from the vehicle coordinate frame) generates safe, feasible trajectories while the global search (illustrated by the four-connected grid) provides guidance toward the goal state. Each motion planning level considers different degrees of fidelity and constraints in order to achieve real-time replanning.

1.2 Constraints

In mobile robot motion planning, constraints are limitations on the inputs or states that a vehicle can achieve. Common types of constraints maintain safety and integrity of the vehicle, such as kinematic and dynamic limits (joint/velocity/torque constraints) and collision avoidance. It is generally important to estimate the response of the suspension/chassis model with the terrain to determine if a vehicle will be subject to tipping over or high centering. Satisfaction of these constraints allow a motion to be considered feasible, or that the path or trajectory is realizable by the vehicle. Computational constraints are also often significant in mobile robot motion planning because resources (processing and memory) are often limited onboard platforms.



■ Fig. 55.1

An example of a hierarchical navigation architecture. A global motion planner (*discrete grid*) provides guidance to a local motion planner (*continuous arc set*) to determine the best trajectory toward the goal state

1.3 Nomenclature

The following notation will be used throughout this chapter:

- Configuration: Vector of independent parameters uniquely specifying the position and orientation of every component of a robotic system relative to a fixed coordinate system. The *configuration space* is the space of all the configurations.
- Configuration Obstacle: Image of a workspace obstacle in the configuration space.
- Path: Continuous sequence of configurations.
- State (\mathbf{x}): Vector of parameters that completely characterize a system in terms of positions, orientations, velocity, etc., at a particular moment in time. The *state space* is the space of all the states.
- Input ($\mathbf{u}(\mathbf{x}, t)$): Control signal that changes the state of the system. May also be referred to as an action or a control.
- Motion Model ($\dot{\mathbf{x}}(\mathbf{x}, \mathbf{u}, t)$): Set of equations that describes the state change as a function of the current state, input, and time.
- Trajectory ($\mathbf{x}(t)$): Continuous sequence of states.
- Space-Time: The space which is obtained by adding the time dimension to a given space.
- State Constraint (\mathbf{x}_C): A vector of constraints that define inclusive (goals, roadways) or exclusive (obstacle) regions in the state space.

2 Motion Safety

Motion safety, i.e., the ability to avoid collisions, is a critical and challenging issue for intelligent vehicles. Accordingly, there is a rich literature on collision avoidance and collision-free motion planning. Before reviewing the most relevant works in [Sect. 2.4](#), the concept of inevitable collision states (ICS) is introduced first in [Sect. 2.1](#). The ICS concept provides insight into collision avoidance and helps in understanding key aspects related to the motion safety issue. These key aspects are analyzed in [Sect. 2.2](#). Then, because motion safety is highly related to the future evolution of the environment, the main approaches to modeling the future are presented in [Sect. 2.3](#).

2.1 Inevitable Collision States

Historically, motion safety boiled down to collision avoidance. The primary goal of the earliest motion planners was to compute geometric paths avoiding collision with the fixed obstacles present in the robot's workspace. In this respect, Lozano-Perez's (1983) configuration space (C-space) is the framework of choice and configuration obstacles (C-obstacles), i.e., the images in the C-space of the workspace's obstacles, capture the no-collision constraints in the form of forbidden regions. Later on, state space (*aka* phase space) became the framework of choice when differential constraints and robots'

dynamics started to be taken into account. In this case, the state obstacles, i.e., the state space counterparts of the C-obstacles, represent the no-collision constraints. Moving obstacles were dealt with in a similar manner thanks to the space-time concept. Adding the time dimension either to the configuration space (Erdmann and Lozano-Perez 1987) or to the state space (Fraichard 1993) allows to model the future evolution of the moving obstacles and again to represent the corresponding no-collision constraints in the form of forbidden regions in the space-time at hand.

In all cases, a configuration or a state in one of the aforementioned regions is forbidden because it yields a collision between the robot and at least one workspace obstacle. However, there is much more to motion safety than mere instantaneous no-collision. Imagine a robotic car traveling very fast toward and a few meters away from a wall. Although the car is not in collision at the present time a collision with the wall is inevitable. Due to momentum, it will crash regardless of any efforts to stop or steer. The concept of *inevitable collision states* (ICS) developed in Fraichard and Asama (2004) can be called upon to account for such a situation. (ICS is similar to the concepts of Obstacle Shadows (Reif and Sharir 1985) or Regions of Inevitable Collision (LaValle and Kuffner 1999). Related concepts are: *Viability Kernels* (Aubin 1991), *Backward Reachable Sets* (Mitchell and Tomlin 2003), and *Barrier Certificates* (Prajna et al. 2007)). An ICS is a state for which, no matter what the future trajectory of the robot is, a collision eventually occurs. The set of ICS defines a region (it is a superset of the set of collision states) in the state space of the robot which, from a motion safety perspective, should be avoided.

A simple scenario illustrating the ICS concept is presented in Fig. 55.2. Imagine a trash compactor or a car crusher, it can be modeled in 2D by two rectangular plates, one of them moving toward the other (Fig. 55.2, left). Let us put a robot \mathcal{A} in the middle of the compactor. To avoid being crushed, \mathcal{A} has to move to the right or to the left until it exits the compactor. To further simplify the problem, \mathcal{A} is treated like a 1D robot that moves along a horizontal line (henceforth called the *position line*). Assuming that \mathcal{A} is

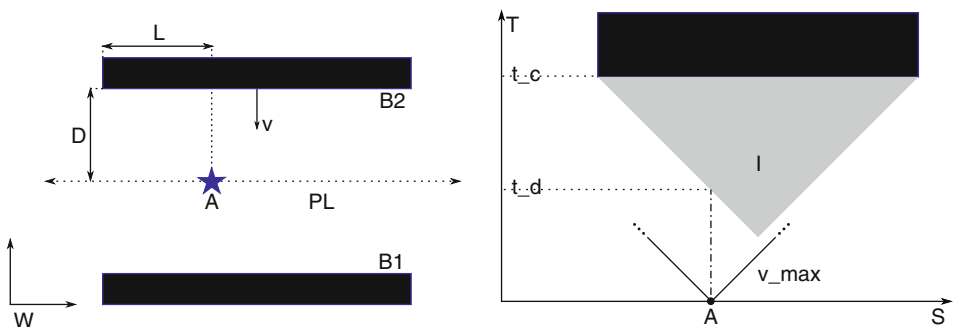


Fig. 55.2

Compactor Scenario (left): the plate B_1 moves toward the plate B_2 . **Corresponding state-time space $X \times T$ (right):** It features both the set of collision states (CS) and inevitable collision states (ICS)

a velocity-controlled point robot, a state of \mathcal{A} is p , i.e., the scalar position of \mathcal{A} on the position line. The motion model of \mathcal{A} is $\dot{p} = v$ with v the upper-bounded velocity of \mathcal{A} ($|v| \leq v_{\max}$). The state-time space of \mathcal{A} consists of two-dimensional position and time (► Fig. 55.2, right). During its motion, \mathcal{B}_1 sweeps across the position line from time $t_c = d_1/v_1$ onward (v_1 is the constant velocity of \mathcal{B}_1 ; d_1 is the distance between \mathcal{A} and \mathcal{B}_1). It yields a set of collision state-times (the black rectangle labeled CS in ► Fig. 55.2, right). CS is a forbidden region that \mathcal{A} must avoid. Likewise, the set of ICS is straightforward to characterize. It is the gray triangular region underneath the CS region (► Fig. 55.2, right). Because of the upper-bound on \mathcal{A} 's velocity, as soon as its state is anywhere inside this gray region, \mathcal{A} is doomed. It will not have the time to escape the compactor. It is worth noting that the ICS definition depends on both the robot (its dynamics) and the environment (the future behavior of the obstacles).

In such a situation, \mathcal{A} might decide to stand still since there is no immediate danger (in fact, no danger before t_c). By doing so, it runs the risk of entering the ICS region and, if that happens, it will be too late. As simple as this example may be, it emphasizes the fact that planning collision-free motions is not enough, it is necessary to consider ICS regions to ensure motion safety. The key to motion safety is therefore to plan motions that stay away from ICS.

2.2 Motion Safety Criteria

The motion safety issue was explored at an abstract level in Fraichard (2007). It laid down three *motion safety criteria* whose violation is likely to put a robotic system into danger and yield collisions. These requirements are fairly intuitive and straightforward to express in two sentences:


- In a dynamic environment, one has a *limited time* only to make a motion decision. One has to *reason about the future* evolution of the environment and do so with an *appropriate lookahead* (The lookahead [aka time horizon] is how far into the future the reasoning is done).

These requirements can be illustrated using the compactor scenario in relation with the ICS concept.

2.2.1 Limited Decision Time

To avoid being crushed, \mathcal{A} has to move right or left until it exits the compactor. Let l_1 denote the distance to the nearest exit (on the left side in this case), the minimum time for \mathcal{A} to escape the compactor is $\delta_e = l_1/v_{\max}$. If \mathcal{A} decides to move to the left, it should start moving before time $t_d = t_c - \delta_e$ which is precisely the time where \mathcal{A} would enter the ICS region if it were standing still. In the situation depicted in ► Fig. 55.2, right t_d represents the upper-bound on the time available of \mathcal{A} to make a motion decision. It is important to note that t_d depends on the current situation and that it can be arbitrarily small (e.g., if \mathcal{B}_1 is larger, then $t_d \rightarrow 0$).

2.2.2 Reasoning About the Future

The space-time model in  Fig. 55.2, right clearly shows that if the future evolution of \mathcal{B}_1 is not taken into account (e.g., if \mathcal{B}_1 is treated like a fixed obstacle), the region CS does not appear in the space-time (neither does the ICS region) and \mathcal{A} cannot be aware of the upcoming collision risk hence the importance of modeling and reasoning about the future evolution of the moving obstacles.

2.2.3 Appropriate Lookahead

Section 2.2.2 has shown the necessity to model the future evolution of the environment and reason about it. The question now is: With what *lookahead*? In other words, how far into the future should the modeling/reasoning go? In the compactor scenario, the answer is straightforward: The lookahead t_{la} must be greater than $t_d + \delta_e$. Otherwise, by the time \mathcal{A} becomes aware of the risk caused by \mathcal{B}_1 , it no longer has the time to decide that it should move to the left and execute this motion. In general, the lookahead must be appropriate, i.e., selected so as to yield a correct definition of the ICS regions. It is important to note that t_{la} depends on the current situation and that it can be arbitrarily large (e.g., if \mathcal{B}_1 is very large and very slow, i.e., $l_1 \rightarrow \infty$ and $v_1 \rightarrow 0$, then $t_{la} \rightarrow \infty$).

In summary, motion safety boils down to three rules (they could be called the *three laws of motion safety*):

1. Decision time constraint: *upper-bounded decision time*.
2. Reasoning about the future: *model of the future required*.
3. Appropriate lookahead: *lower-bounded lookahead*.

The three rules stated above are all related to time. In a dynamic environment, the time dimension is the key aspect. From a motion safety perspective, these requirements should be satisfied otherwise collisions are likely to happen. In a given situation, assuming that a model of the future is available up to the appropriate lookahead, the key to motion safety is to characterize the corresponding ICS regions and to plan a motion that avoids them.

2.3 Future Modeling

Section 2.2 has established the necessity to model the future evolution of the environment up to an appropriate lookahead. Building such a space-time model is in itself a challenge in most real-world situations because complete information about the environment and its future evolution is not available. The purpose of this section is to overview the main approaches that have been proposed in order to address this challenge.

2.3.1 Deterministic Models

The earliest approaches would adhere to a principle that could be dubbed *frozen world*: Every obstacle in the robot's environment is treated as a fixed obstacle and assumed to stay put in the future, e.g., (Khatib 1986a; Borenstein and Korem 1991). Later with the progress of perception in the area of the detection and tracking of moving obstacles, models of the future based on the extrapolation of the moving obstacles' future behavior from their current state appeared. In most cases, the extrapolation relied upon a constant behavior assumption, e.g., constant velocity (Fiorini and Shiller 1998), or regression techniques (Elnagar and Gupta 1998). In a few cases, sophisticated long-term motion prediction techniques have been proposed: They would either exploit the structure of the environment at hand or learn how the obstacles move in a given environment (see Vasquez 2007 and the references therein). In all cases, the common trait of these approaches is that each obstacle is assigned a nominal future motion (known a priori, estimated, or learned). Motion decisions are based upon these nominal future motions.

2.3.2 Conservative Models

From a motion safety perspective, deterministic models are useful as long as their prediction of the future evolution of the environment is reliable. Unfortunately, this reliability can decrease dramatically in the long term. To address this issue, conservative models of the future evolution of the environment have been proposed. The idea is to consider all possible future motions of the environment's obstacles. Accordingly, each obstacle is assigned its reachable set, i.e., the set of states it can reach in the future, to represent its future motion. This is for instance the approach chosen in Schmidt et al. (2006), van den Berg and Overmars (2008).

2.3.3 Probabilistic Models

Conservative models are satisfactory from a motion safety perspective since they can guarantee collision avoidance. However, because of the rapid growth of reachable sets of the obstacles, the whole state space of the robot is eventually forbidden and the robot is blocked. To address this issue, probabilistic models of the future have been proposed. In such models, the future motion of each obstacle is characterized by a probability density function. The tools used to predict the behavior of the moving obstacles are very diverse, e.g., Markov Chains (Rohrmuller et al. 2008), Hidden Markov Models (Vasquez et al. 2009), and Monte Carlo Simulation (Broadhurst et al. 2005). To address motion safety with probabilistic models of the future, Althoff et al. (2010) and Bautin et al. (2010) have proposed probabilistic extension of the ICS concept, probabilistic models are better to capture the uncertainty that prevails in the real world, in particular the uncertainty concerning the

future behavior of the moving obstacles. However, they do not allow strict motion safety guarantee, they allow instead to minimize the risk.

2.4 Literature Review

Section 2.1 has established the fact that, broadly speaking, motion safety for a given robot can be achieved by staying away from the ICS regions. The challenge then for a robot is to characterize the ICS set corresponding to its current situation and to move so as to avoid it. Although the ICS concept has been around for at least 20 years now (under other names sometimes), it is only recently that it has surfaced as a key aspect of motion safety (See the recent ICRA workshop on this topic: <http://safety2010.inrialpes.fr>). Reviewing the vast literature on collision avoidance and safe navigation, it is possible to classify the related works in different families corresponding to the answers that are put forward in order to obtain motion safety. These families are presented in the next sections.

Reactivity. These approaches rely on the implicit assumption that the ability to react in real time to unexpected events is sufficient to guarantee collision avoidance. Accordingly, they focus on developing decision-making schemes that are fast. They usually adhere to the frozen-world principle. Approaches of this family are the earliest in the field, e.g., (Khatib 1986a; Fox et al. 1997; Minguez and Montano 2004; Ulrich and Borenstein 2000), although recent autonomous vehicles still follow this line of approach, e.g., (Xia et al. 2010). In the light of the ICS concept, there is no need to point out that such approaches are not satisfactory for motion safety in dynamic environments.

τ -safety. Unlike the previous family of methods, these approaches do reason about the future behavior of the moving obstacles. The future motion of the robot is computed so as to guarantee that the robot can safely reach states (usually equilibrium states) wherein it can stay safely for a duration τ (hopefully sufficient to compute an updated safe trajectory at the next planning cycle), e.g., (Frazzoli et al. 2002; Vatcha and Xiao 2008). Once again, in the light of the ICS concept, the τ -safety condition may not be enough to guarantee motion safety (an ICS state can be τ -safe).

ICS approximation. Given the complexity of the ICS set characterization, a number of approaches rely upon ICS approximations. Such approximations can be obtained by (1) considering the moving obstacles independently and (2) considering an arbitrary lookahead, and even using learning, e.g., (Chan et al. 2007; Kalisiak and van de Panne 2007). These approximations being not conservative, there is always a risk that a state labeled as safe is in fact an ICS.

Evasive trajectories. These approaches are by far the most interesting. They acknowledge the difficulty of getting a meaningful characterization of the ICS set in real-world situation. Such a characterization would be the key to absolute motion safety, i.e., the guarantee that collision could be avoided no matter what happens. Instead, these approaches rely upon a relaxation of the ICS concept: They seek to guarantee that the robot can only be in states where it is possible to execute an evasive trajectory,

e.g., a braking maneuver for a car or a circling maneuver for a plane. Examples for these kinds of approaches can be found in Hsu et al. (2002), Petti and Fraichard (2005); Bekris and Kavraki (2007); Seder and Petrovic (2007).

3 Iterative Motion Planning

The previous section reviewed methods for guaranteeing or approximating the safety of a trajectory. This section discusses how to efficiently sample the space of trajectories. Iterative motion planning algorithms are generally distinct from others in robotics because efficiency and real-time performance is often favored over true optimality. The main differences between different types of iterative motion planning algorithms involve the way that the space of feasible motions are represented, approximated, or sampled. This section focuses on the local motion planning component of hierarchical navigation and discusses potential field (➤ Sect. 3.1), sampling (➤ Sect. 3.2), and graph search techniques (➤ Sect. 3.3) to achieve real-time iterative motion planning for mobile robot navigation.

3.1 Potential Field Techniques

Motion planning using artificial potential fields (Khatib 1986b) have been applied to the problem of mobile robot navigation (Haddad et al. 1998) by representing goals and obstacles as attractive and repulsive forces respectively. In Haddad et al. (1998), goal potentials are evaluated as quadratic functions of goal distance and obstacle potentials are computed as a linear combination of position and heading proximity. Potential field techniques compute the resulting potential as the sum of goal and obstacles and use the gradient to generate a control in the direction of steepest descent. In Iagnemma et al. (2008), potential fields are extended to the curvature-velocity space and are composed of the sum of rollover, sideslip, waypoint, and hazard potentials. This approach guarantees safety of the generated trajectory by post-processing the resulting motion and reducing the commanded velocity of any inputs that violate safety constraints.

3.2 Sampling Techniques

Sampling techniques for iterative motion planning are generally based on generating and sorting a set of candidate motions. The fundamental challenge of sampling-based techniques involves determining how best to efficiently sample the space of feasible motions. This section will review research in input space sampling (➤ Sect. 3.2.1) and state space sampling (➤ Sect. 3.2.2) techniques with application to mobile robot motion navigation.

3.2.1 Input Space Sampling

Some of the earliest work in producing feasible iterative motion planning search spaces through input space sampling appears in Kelly and Stentz (1998). This approach differs from prior techniques that sampled in the space of geometric paths because each input is simulated with a vehicle model to determine the state space response. Input space sampling is a straightforward technique for generating a feasible motion planning search space because all motions sampled in the input space are inherently executable. This approach has been successfully applied in a number of mobile robot applications including high-speed field robots (Kelly et al. 2006) and explore terrestrial (Wettergreen et al. 2005) and extraterrestrial environments (Biesiadecki and Maimone 2006). Algorithm 1 presents an overview of the input space sampling search space generation technique. Each input $\mathbf{u}_i(\mathbf{x}, t)$ is simulated using the predictive motion model $\dot{\mathbf{x}}(\mathbf{u}, \mathbf{x}, t)$ to determine the shape of the resulting trajectory. That trajectory is then evaluated for safety using methods including those discussed in the previous section.

Algorithm 1: Input Space Sampling Based Search Space Generation

input : $\mathbf{x}(t_i)$, $\dot{\mathbf{x}}(\mathbf{u}, \mathbf{x}, t)$

output: $\mathbf{x}_N(t)$, $\mathbf{u}_N(\mathbf{x}, t)$, c_N

begin

```

  |  $\mathbf{u}_N(\mathbf{x}, t) \leftarrow \text{SAMPLEINPUTSPACE}(\mathbf{x}(t_i));$ 
  | for  $i \leftarrow 1$  to  $N$  do
  |   |  $\mathbf{x}_i(t) \leftarrow \text{SIMULATEACTION}(\mathbf{x}(t_i), \mathbf{u}_i(\mathbf{x}, t), \dot{\mathbf{x}}(\mathbf{u}, \mathbf{x}, t), \mathbf{x}, t);$ 
  |   |  $c_i \leftarrow \text{COMPUTETRAJECTORYCOST}(\mathbf{x}_i(t));$ 
  | end

```

end

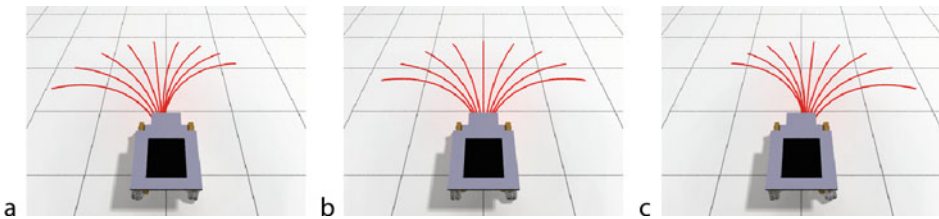
The input space sampling technique applies some logic or function to determine the set of actions to simulate. A predictive model evaluates the state space response of the action from the initial state and a utility or cost function determines the quality of the resulting motion. The result is a trajectory set (\mathbf{x}_N) of actions $\mathbf{u}(\mathbf{x}, t)$ that can be sorted for maximum utility or minimum cost (c) to execute. An example of motion set generated using the input space sampling technique is illustrated in [Fig. 55.3](#).

The example in [Fig. 55.3](#) shows the space response of a uniformly sampled input space subject to different initial conditions. The search spaces shown in [Figs. 55.3a, b, and c](#) are all generated by simulating nine constant curvature inputs sampled between $-0.8 \frac{\text{rad}}{\text{m}} \leq k \leq 0.8 \frac{\text{rad}}{\text{m}}$ for a fixed duration. The differences in initial state and the constraints of the vehicle model can cause many samples in the input space to produce similar state space trajectories as evident in [Fig. 55.3a, c](#). Since these overlapping trajectories exhibit very little separation, evaluating these trajectory sets are not as efficient as others with motions that are more expressive.

Input space sampling techniques can be applied to search in the spaces of actions that are more expressive than constant curvature arcs. [Figure 55.4a](#) shows a search space produced by sampling in the space of curvature clothoids. These nine trajectories are

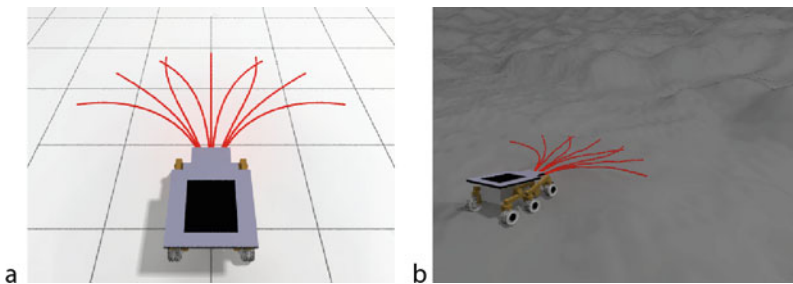
generally more diverse than those in [Fig. 55.3](#) because the space of reachable headings is better represented. Trajectory sets computed using input space sampling techniques can also be easily adapted for use on rough terrain by simulating the interaction between the vehicle chassis and the terrain during each action. Since the technique only requires that each action can be simulated, an arbitrarily complex predictive motion model can be applied to generate the trajectory set. The resulting search space ([Fig. 55.4b](#)) is now more informed about the response of the vehicle chassis at each point along the sampled input without modifying the search space generation technique.

Generation of efficient input space sampling search spaces is still an active research topic. Knepper et al. (2010) have investigated removing homotopically identical trajectories from trajectory sets to reduce the number of required collision detection computations. There has also been complimentary work in online identification of predictive motion models to improve the fidelity of resulting search space (Rogers-Marcovitz and Kelly 2010).



■ Fig. 55.3

Trajectory sets generated by input space sampling with varying initial state. Each body-frame action is defined as a constant function of curvature and linear velocity. (a) Sampling in the state space with a large negative initial curvature. (b) Sampling in the state space with a zero initial curvature. (c) Sampling in the state space with a large positive initial curvature



■ Fig. 55.4

Variations of input space sampling techniques including sampling in higher dimensional input spaces and on rough terrain. (a) Sampling in the input space of clothoids with a zero initial curvature. (b) Sampling in the input space of clothoids on with zero initial curvature on a rough terrain

3.2.2 State Space Sampling

An alternative to input space sampling techniques involves sampling in the state space of vehicle motion. State space sampling provides more control over the distribution and expressiveness of motions in the trajectory set by specifying the boundary conditions of each motion. The trade-off of this approach is that feasibility is no longer guaranteed – an inverse trajectory generation method must be used in order to determine what (if any) input can satisfy the boundary state constraints and produce the motion (Howard et al. 2008). Algorithm 2 presents an overview of the state space sampling search space generation technique.

Algorithm 2: State Space Sampling Based Search Space Generation

```

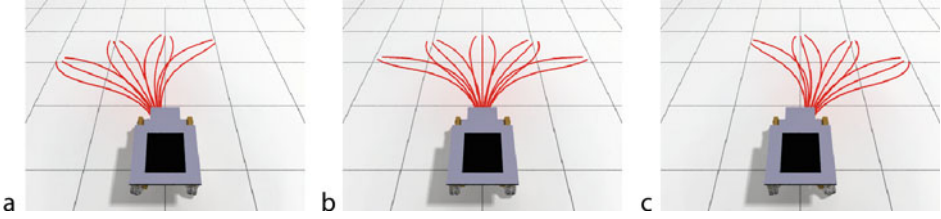
input :  $\mathbf{x}(t_i)$ ,  $\dot{\mathbf{x}}(\mathbf{u}, \mathbf{x}, t)$ 
output:  $\mathbf{x}_N(t)$ ,  $\mathbf{u}_N(\mathbf{x}, t)$ ,  $c_N$ 
begin
  |  $\mathbf{x}_C \leftarrow \text{GENERATEBOUNDARYSTATES}();$ 
  | for  $i \leftarrow 1$  to  $N$  do
  |   |  $[\mathbf{u}(\mathbf{x}, t), \mathbf{x}_i(t)] \leftarrow \text{GENERATETRAJECTORY}(\mathbf{x}(t_i), \dot{\mathbf{x}}(\mathbf{u}, \mathbf{x}, t), \mathbf{x}_C);$ 
  |   | if  $\mathbf{u}(\mathbf{x}, t)$  exists then
  |   |   |  $c_i \leftarrow \text{COMPUTETRAJECTORYCOST}(\mathbf{x}_i(t));$ 
  |   |   else
  |   |   |  $c_i \leftarrow \infty;$ 
  |   |   end
  |   end
end

```

There are two significant differences between Algorithms 1 and 2. The first difference is that the state space sampling algorithm requires that the inputs must be solved for each constraint pair in the boundary state set online. This is the inverse trajectory generation problem where for a mobile robot at a particular initial state ($\mathbf{x}(t_i)$) an action ($\mathbf{u}(\mathbf{x}, t)$) must be found that satisfies both the motion model ($\dot{\mathbf{x}}(\mathbf{x}, \mathbf{u}, t)$) and the boundary state constraints (\mathbf{x}_C). An efficient solution to this problem is necessary because it must be solved many times online. The second difference involves developing a method for sampling in the space of boundary state constraints instead of input space.

The main difference between Algorithms 1 and 2 is that the state space sampling algorithm requires that trajectories be found to satisfy the boundary state pairs at runtime. This is the inverse trajectory generation problem where an input ($\mathbf{u}(\mathbf{x}, t)$) must be found that satisfies the initial state ($\mathbf{x}(t_i)$), motion model ($\dot{\mathbf{x}}(\mathbf{x}, \mathbf{u}, t)$), and boundary state (\mathbf{x}_C) constraints. This technique must be efficient because it is computed frequently during search space synthesis. A view of trajectory sets constructed using the state space sampling technique with varying initial state is shown in [Fig. 55.5](#).

The trajectory sets illustrated in [Fig. 55.5](#) differ from those in [Fig. 55.3](#) or [Fig. 55.4](#) because the shape of the search space is much less dependent on the initial conditions of the robot. Each motion in the trajectory set is computed by determining an input that satisfies initial state, motion model, and sampled terminal position and heading



■ Fig. 55.5

Trajectory sets generated by state space sampling with varying initial state. Each body-frame action is defined as a quadratic function of curvature and constant function of linear velocity with sampled terminal position and heading boundary state constraints at a constant horizon. (a) A state space sampling search space with a nearby obstacle. (b) Biasing the search space away from a front facing obstacle. (c) Biasing the search space away from an off-center obstacle

constraints at a fixed horizon. The space sampling techniques are able to expressively sample the state space of vehicle motion using inputs based on a quadratic function of curvature and a constant function of linear velocity.

Model-Predictive Trajectory Generation. Many techniques have been proposed through the years to solve the two-point boundary value mobile robot trajectory generation problem including clothoids (Kanayama and Miyake 1985; Shin and Singh 1991), cubic polynomials (Kanayama and Hartman 1989), shooting methods (Jackson and Crouch 1991), energy minimization (Delingette et al. 1991), bang-bang control (Kalmár-Nagy et al. 2004), sinusoidal and Fourier series (Brockett 1981; Tilbury et al. 1992; Murray and Sastry 1993), Bézier splines (Komoriya and Tanie 1989; Shiller and Chen 1990), and sequences of cubic splines (Amar et al. 1993). One approach (Kelly and Nagy 2003; Howard and Kelly 2007) involves parameterizing the space of inputs to reduce the space of possible motions:

$$\mathbf{u}(\mathbf{x}, t) \rightarrow \mathbf{u}(\mathbf{p}, \mathbf{x}, t) \quad (55.1)$$

With this reduced search space, the optimal control problem can be converted to parametric optimal control that can be solved in real time. From an initial guess of parameters \mathbf{p}_0 , an iterative correction can be found by estimating the Jacobian of terminal state constraint error with respect to parameterized freedom perturbations:

$$\mathbf{x}_F(\mathbf{p}) = \mathbf{x}(t_f) + \int_{t_i}^{t_f} \dot{\mathbf{x}}(\mathbf{u}(\mathbf{p}, \mathbf{x}, t), \mathbf{x}, t) dt \quad (55.2)$$

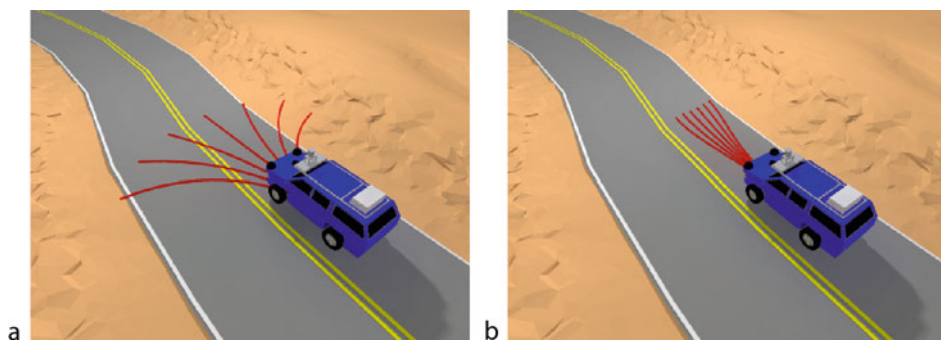
$$\mathbf{p}_{i+1} = \mathbf{p}_i - \left[\frac{\partial(\mathbf{x}_F(\mathbf{p}) - \mathbf{x}_C)}{\partial \mathbf{p}} \right]^{-1} [\mathbf{x}_F(\mathbf{p}) - \mathbf{x}_C], \quad 0 \leq i \leq n \quad (55.3)$$

Often partial derivatives in the Jacobian cannot be computed analytically for vehicles with non-holonomic constraints or in rough terrain. Numerical techniques (forward or central differences) have proven to be an effective, model-invariant approach.

Two significant challenges with parametric optimal control methods include parameter initialization and computational complexity. The convergence performance of the algorithm can be greatly improved if the initial guess of parameters is near the continuum solution. Precomputed parameter lookup tables based on flat-terrain approximations of solutions (Howard 2009) have demonstrated to be one effective method of parameter initialization for model-predictive trajectory generation. The computational complexity of the algorithm is related to the quality of the initial parameter guess because runtime is approximately linear with respect to the number of iterations required to achieve the desired accuracy. Trading off predictive motion model fidelity for the computational efficiency is a technique to scale the runtime of the algorithm to a particular application.

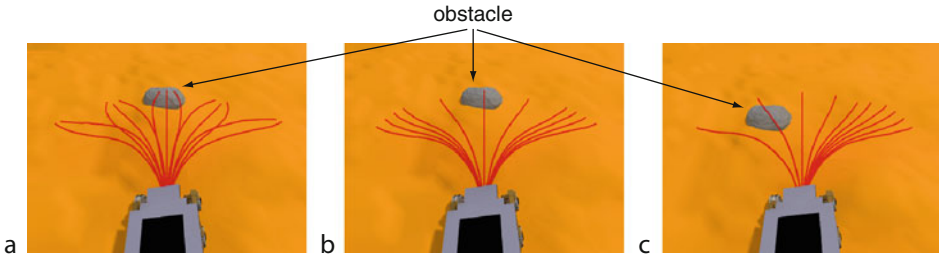
Boundary State Sampling Techniques. One advantage of the state space sampling technique is that the sampling function can be continuously adapted to fit the application or environment. Ultimately, the design of the sampling function is driven by the constraints of the vehicle and the environment. In a constrained environment such as a road or trail, it is beneficial to sample only in the terminal state space inside the corridor or lane. The computational cost of evaluating trajectory safety for motions known to terminate outside of safe regions can be reduced by sampling in the terminal state-space of obstacle-free collisions. An example of this approach is illustrated in [Fig. 55.6](#).

In the absence of environmental constraints, expressiveness and efficiency become the focus of the state space sampling design method. It has already been shown in [Fig. 55.5](#) that state space sampled search space can produce a trajectory set that is more uniformly distributed in state space than one sampled in the input space; however, an interesting extension of this technique involves biasing the samples toward the minimum global cost regions ([Fig. 55.7](#)) (Howard et al. 2008). In this example, the responses of three state space sampled search spaces are shown in the presence of nearby obstacle. In [Fig. 55.7a](#),



■ Fig. 55.6

Efficient sampling in constrained environments. Path sets can be constructed using boundary states determined by geometry in the environment such as roads or trails to increase the likelihood of finding a collision-free path. (a) A search space generated using the input space sampling technique in a constrained environment. (b) A search space generated using the state space sampling technique in a constrained environment



■ Fig. 55.7

Efficient sampling in unconstrained environments. Path sets with fewer collisions can be determined by biasing the state space samples toward regions with minimum global motion planning cost

the search space is uniformly distributed and many of the solutions collide with the front facing rock. If the global motion planner is used to bias the samples toward regions with minimum global cost (► Fig. 55.7b, c), many more of the paths in the trajectory sets do not collide with the obstacle.

State space sampling techniques are most effective when there is structure to the environment although the algorithm has the ability to exploit other types of global guidance to generate efficient trajectory sets. The ability to bias the search space toward regions of interest allows sampling-based techniques be an effective solution for local motion planning in a hierarchical motion planning architecture.

3.3 Graph Search Techniques

Graph search techniques are a step beyond input or state space sampling techniques toward more informed motion planners. As the complexity of primitive motions in sampling-based motion planner increases (i.e., forward/reverse actions, multipoint turns), graph search techniques provide a more efficient way to search the space of feasible motions. Graph search composes an efficient search space by expanding only in desirable regions. Derivative motions from edges that intersect with obstacles are never evaluated because any of those trajectories would be deemed hazardous. This section describes several approaches including ego-graphs (► Sect. 3.3.1), rapidly exploring dense trees (► Sect. 3.3.2) and recombinant state lattices (► Sect. 3.3.3) for mobile robot navigation.

3.3.1 Ego-graphs

Ego-graph-based iterative motion planning represents early work in applying graph search with feasible motion planning search spaces. This method combined off-line and online algorithms to achieve real-time replanning in unstructured outdoor environments. The off-line portion generated dynamically feasible edges between five layers of 17

sampled poses relative in a body-fixed reference frame. Since the shape of the search space is highly dependent on the initial steering angle and vehicle velocity, many sets are generated and stored by sampling in the space of these values. The online portion of this algorithm searches this vehicle-centric tree using A* graph search (Hart et al. 1968) to determine a safe, feasible local motion plan. While this technique in its original form did not compensate for disturbances including terrain shape and degrading mobility, real-time model-predictive trajectory generation techniques could be applied to refine the search space and adapt the first level of motions to the continuous initial velocity and steering angle.

3.3.2 Rapidly Exploring Dense Trees

Algorithm 3: RRT Building

```

 $\mathcal{T} \leftarrow \mathbf{x}_{\text{start}};$ 
for  $i = 1$  to  $n_{\text{steps}}$  do
    |  $\mathbf{x}_{\text{rand}} \leftarrow \text{RANDOM\_STATE}();$ 
    |  $\text{EXTEND}(\mathcal{T}, \mathbf{x}_{\text{rand}});$ 
end

```

The Rapidly Exploring Dense Tree (RDT) family of methods belongs to the single-query sampling-based motion planner category (Lavalle 2006, ► Chap. 5). The idea is to incrementally build a search tree until the goal is reached. The incremental growth of the tree relies upon a dense sequence of samples. In the limit, as the number of samples tends to infinity, the tree densely covers the search space. This denseness ensures probabilistic completeness. Note that RDTs avoid maintaining a lattice.

The most popular RDT algorithm is the Rapidly Exploring Random Tree (RRT) method introduced in (Lavalle 1998), it uses a random sequence of samples and biases the exploration toward the unexplored region of the search space. The basic RRT construction algorithm is outlined in Algorithm 3. Each cycle of Algorithm 3 attempts to extend the RRT by adding a new vertex which is biased by a randomly selected state \mathbf{x}_{rand} . The *EXTEND* function selects the RRT vertex \mathbf{x}_{near} which is closest to \mathbf{x}_{rand} and computes a new vertex \mathbf{x}_{new} by moving toward \mathbf{x}_{rand} for a given incremental time/distance, e.g., by applying a constant control over a given time interval. If the trajectory between \mathbf{x}_{near} and \mathbf{x}_{new} is collision free then the corresponding edge is added to the RRT. It is straightforward to design a motion planner using Algorithm 3 by growing a tree \mathcal{T} rooted at the start state $\mathbf{x}_{\text{start}}$ and periodically checking whether it is possible to connect \mathcal{T} to the goal state \mathbf{x}_{goal} . To that end, it suffices that the *EXTEND* function periodically returns \mathbf{x}_{goal} .

RRT-based motion planning is particularly suited to address problems with differential constraints. Besides implementation technicalities, e.g., nearest vertex computation, its main drawback is the sensitivity of its performance on the choice of the metric which is used to specify the distance between pairs of states (Cheng and LaValle 2001). In spite of this, RRT is perhaps the most popular single-query motion planner with new applications

and new extensions presented every year. RDTs in general and RRTs in particular have displayed excellent performance for a variety of motion planning problems.

In the iterative motion planning context, RRT can be used to grow a tree toward the next waypoint or waystate. When the planning time is over, the best trajectory (according to a given criterion) is extracted from the tree and returned.

3.3.3 Recombinant State Lattices

Recombinant state lattices (Ferguson et al. 2008b; Pivtoraiko et al. 2009) represent another step toward near-optimal motion planning algorithms that are both feasible and efficient. A state lattice is some specialized discretization of the state space designed to represent the mobility of an autonomous vehicle. It generalizes the idea of a four- or eight-connected grid, Dubins car (1957), and Reeds–Shepp car (1990) to a precomputed set of edges (*control set*) that describes the connectivity of the motion planning graph. The technique generates a search space by repeatedly expanding and evaluating control sets in this multidimensional state lattice uses heuristic graph search algorithms A^* , D^* (Stentz and Herbert (1995), or D^* Lite (Koenig and Likhachev 2002) to determine the minimum-cost path.

A variation of this technique introduces the concept of *graduated fidelity* which varies the expressiveness of the control set with the proximity to the current state. Similar to other multi-resolution techniques (Ferguson and Stentz 2006), it achieves fast performance by searching densely in regions where perceptual information is available. This is particularly well suited for hierarchical navigation architectures because it combines the ideas of local and global motion planners into a single unified technique.

4 Applications

Iterative motion planners have demonstrated the capability to safely navigate in urban, off-road, and even extraterrestrial environments. This section reviews the research in on-road (● Sect. 4.1) and off-road (● Sect. 4.2) mobile robots with respect to iterative motion planning and safety.

4.1 Roadway and Urban Navigation

The first known attempt to build an autonomous roadway vehicle was in 1977. A car developed by Tsukuba Mechanical Engineering Laboratory in Japan was able to follow white street markers and to reach speeds of up to 30 km/h on a dedicated test course.

A few years later, from 1987 to 1995, Ernst Dickmanns developed the VaMoRs and VaMoRs-P vehicles (*aka* VITA and VITA-II). The first one, a Mercedes-Benz D811 van Ulmer (1992) achieved 100 km/h on streets without traffic. The second one,

a Mercedes-Benz S-Class SEL500 (Ulmer 1994), made extended trips on highways in standard heavy traffic at speeds up to 175 km/h. They both used vision and demonstrated autonomous driving in free lanes, convoy driving, and lane changes with autonomous overtaking. Collision avoidance in VITA-II was handled using an artificial potential field approach (Khatib 1986a). The potential field used was a combination of fields associated with the desired velocity, the moving obstacles around, the lanes and traffic regulations (Reichardt and Shick 1994).

In 1995 and 1999 respectively, two similar events took place on both sides of the Atlantic. The CMU NAVLAB 5, a 1990 Pontiac Trans Sport, completed its 4,600 km “No Hands Across America” tour Jochem et al. (1995). A few years later, in 1999, the ARGO vehicle, a modified Lancia Thema, completed its 2,000 km “MilleMiglia in Automatic” journey across Italy (Broggi et al. 1999). Both vehicles used vision to determine the location of the road ahead and the appropriate steering direction to keep the vehicle on the road (throttle and brakes were handled by a human driver). While obstacle avoidance was not addressed in NAVLAB, ARGO was able to detect obstacles ahead and change lane if need be. Control of the steering angle is done thanks to a feedback controller following a virtual target.

These pioneering attempts at autonomous driving were somehow restricted to highly structured and somewhat simple environments, i.e., highways. The focus was on lane detection using vision and autonomous driving was pretty much reduced to lane following with the odd lane change. Over the last decade things started to change with a European research program called Cybercars (<http://www.cybercars.org>) whose long-term goal was to develop road vehicles with fully automated driving capabilities. A fleet of such vehicles would form a managed transportation system, for passengers or goods, on a network of roads with on-demand and door-to-door capability. In this framework, Benenson et al. (2006) developed a Cybercar prototype called Cycab, a small electric vehicle equipped with range and visual sensors able to detect static and moving obstacles. The navigation scheme proposed and dubbed partial motion planning (Petti and Fraichard 2005) is iterative and graph search-based (RRT): It takes into account both the estimated future behavior of the moving objects and the limits of the sensors’ field of view. Motion safety is enforced through braking maneuvers. Autonomous driving was demonstrated on a small scale in environments featuring pedestrians and other vehicles.

At the same time, the SmartTer (<http://www.smart-team.ch>) platform based on a Smart Fortwo vehicle was demonstrated at the ELROB 2006 event (<http://www.elrob.org>). The noticeable thing about the SmartTer is that it was able to do 3D mapping using rotating LIDAR. Navigation on the SmartTer adheres to the frozen-world assumption and uses input space sampling to determine the local trajectory to follow.

The DARPA Urban Challenge was a autonomous road race held at an former United States Air Force Base in 2007. Six of the eleven finalist finished the 60 mile urban course amongst human and robot-operated vehicles. It was a particular challenge from a motion

safety point of view since each autonomous vehicle had to make “intelligent” decisions in real time based on the actions of the other vehicles. Looking at the four entries that managed to complete the course in the allotted time, it is interesting to note the similarities between their navigation schemes: they are all hierarchical and behavior-based. The structure of the roadway would be used both to estimate the future behavior of the moving obstacles and to determine the future behavior of the vehicle, e.g., lane following. Differences appear in the way local trajectories are generated and how motion safety is addressed:

- CMU’s Boss (Ferguson et al. 2008a) uses state space sampling to generate a set of local trajectories which are then checked for collision with the moving obstacles using hierarchical space-time collision checking.
- Stanford’s Junior (Montemerlo et al. 2008) uses input space sampling and graph search (hybrid A*) for local trajectory computation. Interactions with the moving obstacles are addressed at the behavioral level.
- Virginia Tech’s Odin (Bacha et al. 2008) uses input space sampling and graph search (A*) for local trajectory computation. Motion safety is handled in two stages: a “speed limiter” considers the moving obstacles to generate speed constraints that are then passed to the trajectory generator itself that operates on a static cost-map.
- For MIT’s Talos (Kuwata et al. 2009), trajectories are computed using graph search (RRT-like). Their safety is evaluated based on the availability of a collision-free braking trajectory (assuming that the moving obstacles maintain their current driving behavior).

The development of autonomous vehicles is still going on, e.g., Jia Tong University’s CyberC3 vehicles (Xia et al. 2010), the Vislab Intercontinental Autonomous Challenge (Four electric vans equipped to drive in leader–follower configuration successfully completed a 13,000 km drive from Italy to China in 3 months). Broggi et al. (2010), or Google’s driverless cars Thrun (2010).

4.2 Off-Road Navigation

Autonomous navigation in off-road environments has been an active area of research for more than two decades. Some of the earliest work in unstructured off-road navigation appears in Daily et al. (1988) where the autonomous land vehicle (ALV) demonstrated the ability to drive through natural terrains using real-time vehicle control and a map-based planner. Estimates of suspension response, vehicle attitude, and terrain clearance were used to classify whether the terrain posed an obstacle in the map-based planner. Stentz and Herbert (1995) demonstrated early hierarchical navigation with an arc-based local planner (SMARTY) and an optimal global path planner (D*) on NAVLAB. RANGER (Kelly and Stentz 1998), which was later integrated and demonstrated on the PerceptOR program (Stentz et al. 2002), used a predictive model to estimate the state response of sampled actions in the local motion planner.

The 2004 and 2005 DARPA Grand Challenges were significant events in the development of off-road autonomous robots. The events pitted self-driving vehicles (including modified cars, dune buggies, trucks, and a motorcycle) in a timed race through the Mojave desert. A wide range of techniques for mobile robot navigation were explored and applied on vehicle ranging from a modified motorcycle (Song et al. 2007) to a 13 ton truck (Braid et al. 2006). Many of the techniques in this race applied hierarchical navigators that utilized sampling or graph search techniques to determine obstacle-free paths in the environment. Urmson et al. (2006) searched a trail-oriented graph composed of straight-line edges and relied on a speed planner and feedback control based the pure-pursuit technique Coulter (1992) to follow the paths. Thrun et al. (2006) utilized a search space based on lateral offsets from a fixed-based trajectory and selected actions using a weighted cost function based on maximum lateral acceleration, maximum steering angle, maximum steering rate, and maximum deceleration. Each motion was simulated with a vehicle model to ensure that each path was feasible and local planner lookahead varied based on the vehicle speed.

5 Conclusions and Further Reading

There are many approaches to iterative motion planning and the field is continuously evolving. Even though mobile robots have demonstrated the capability to autonomously navigate through urban, off-road, and even extraterrestrial environment, new techniques are actively research and deployed. Current research in iterative motion planning techniques still seek to most efficiently represent the space of feasible motions in a computationally efficient manner. As mobile robots move faster through rougher terrain (and as representations between free space and obstacle blur) it will become more important to understand the complex interaction between the vehicle and the environment. From the motion safety point of view, it is important to emphasize that, in spite of the advances made and the success of events such as the DARPA Challenges or the Intercontinental Autonomous Challenge, autonomous driving in a full-fledged urban environments, e.g., with pedestrians or cyclists, remains an open problem (consider the accidents that took place during the DARPA Urban Challenge Fletcher et al. (2008)). Motion safety in such urban scenarios requires further advances in both situation analysis, i.e., what is going on now and what will happen next, and decision making.

Acknowledgments

A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

- Althoff D, Althoff M, Wollherr D, Buss M (2010) Probabilistic collision state checker for crowded environments. In: IEEE international conference on robotics and automation, Anchorage. doi:10.1109/ROBOT.2010.5509369
- Amar F, Bidaud P, Oueddou F (1993) On modeling and motion planning of planetary vehicles. In: Proceedings of the 1993 IEEE/RSJ international conference on intelligent robots and systems, vol 2, Yokohama, pp 1381–1386
- Aubin JP (1991) Viability theory. Birkhuser, Boston
- Bacha A, Bauman C, Faruque R, Fleming M, Terwelp C, Reinholtz C, Hong D, Wicks A, Alberi T, Anderson D, Cacciola S, Currier P, Dalton A, Farmer J, Hurdus J, Kimmel S, King P, Taylor A, van Covern D, Webster M (2008) Odin: team VictorTango's entry in the DARPA urban challenge. *J Field Robot* 25(8). doi:10.1002/rob.20248
- Bautin A, Martinez-Gomez L, Fraichard T (2010) Inevitable collision states, a probabilistic perspective. In: IEEE international conference on robotics and automation, Anchorage. doi:10.1109/ROBOT.2010.5509233
- Bekris K, Kavraki L (2007) Greedy but safe replanning under kinodynamic constraints. In: IEEE international conference on robotics and automation, Rome. doi:10.1109/ROBOT.2007.363069
- Benenson R, Petti S, Fraichard T, Parent M (2006) Integrating perception and planning for autonomous navigation of urban vehicles. In: IEEE/RSJ international conference on intelligent robots and systems, Beijing. doi:10.1109/IROS.2006.281806
- Biesiadecki J, Maimone M (2006) The mars exploration rover surface mobility flight software: Driving ambition. In: Proceedings of the 2006 IEEE aerospace conference, Pasadena, 2006
- Borenstein J, Korem Y (1991) The vector field histogram - fast obstacle avoidance for mobile robots. *IEEE Trans Robot Autom* 7(3). doi:10.1109/70.88137
- Braid D, Broggie A, Schmiedel G (2006) The terramax autonomous vehicle. *J Field Robot* 23(9):693–708
- Broadhurst A, Baker S, Kanade T (2005) Monte Carlo road safety reasoning. In: IEEE intelligent vehicles symposium, Las Vegas. doi:10.1109/IVS.2005.1505122
- Brockett R (1981) Control theory and singular Riemann Geometry. Springer, New York
- Broggi A, Bertozzi M, Fascioli A, Guarino Lo Bianco C, Piazzi A (1999) The ARGO autonomous vehicle's vision and control systems. *Int J Intell Contr Syst* 3(4):409–441
- Broggi A, Medici P, Cardarelli E, Cerri P, Giacomazzo A, Finardi N (2010) Development of the control system for the vislab intercontinental autonomous challenge. In: IEEE international conference on intelligent transportation systems, Madeira. doi:10.1109/ITSC.2010.5625001
- Chan N, Zucker M, Kuffner J (2007) Towards safe motion planning for dynamic systems using regions of inevitable collision. In: Collision-free motion planning for dynamic systems workshop, Rome
- Cheng P, LaValle S (2001) Reducing metric sensitivity in randomized trajectory design. In: IEEE/RSJ international conference on intelligent robots and systems, Hawaii. doi:10.1109/IROS.2001.973334
- Coulter R (1992) Implementation of the pure pursuit path tracking algorithm. Technical report, Carnegie Mellon University
- Daily M, Harris J, Keirsey D, Olin K, Payton D, Reiser K, Rosenblatt J, Tseng D, Wong V (1988) Autonomous cross-country navigation with the alv. In: Proceedings of the IEEE international conference on robotics and automation, Philadelphia, pp 718–726
- Delingette H, Gerbert M, Ikeuchi K (1991) Trajectory generation with curvature constraint based on energy minimization. In: Proceedings of the 1991 IEEE/RSJ International Conference on Intelligent Robots and Systems, Osaka, vol 1, pp 206–211
- Dubins L (1957) On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *Am J Math* 79:497–516
- Elnagar A, Gupta K (1998) Motion prediction of moving objects based on autoregressive model. *IEEE Trans Syst Man Cybern A Syst Hum* 28(6). doi:10.1109/3468.725351

- Erdmann M, Lozano-Perez T (1987) On multiple moving objects. *Algorithmica* 2:477–521
- Ferguson D, Stentz A (2006) Multi-resolution field d*. In: *Proceedings of the international conference on intelligent autonomous systems*, Pittsburgh, pp 65–74
- Ferguson D, Darms M, Urmson C, Kolski S (2008a) Detection, prediction, and avoidance of dynamic obstacles in urban environments. In: *IEEE intelligent vehicles symposium*, Eindhoven. doi:10.1109/IVS.2008.4621214
- Ferguson D, Howard T, Likhachev M (2008b) Motion planning in urban environments: part i. In: *Proceedings of the 2008 IEEE/RSJ international conference on intelligent robots and systems*, Hoboken, 2008
- Fiorini P, Shiller Z (1998) Motion planning in dynamic environments using velocity obstacles. *Int J Robot Res* 17(7):760–772
- Fletcher L, Teller S, Olson E, Moore D, Kuwata Y, How J, Leonard J, Miller I, Campbell M, Huttenlocher D, Nathan A, Kline FR (2008) The MIT – cornell collision and why it happened. *Int J Field Robot* 25(10):775–807
- Fox D, Burgard W, Thrun S (1997) The dynamic window approach to collision avoidance. *IEEE Robot Autom Mag* 4(1):23–33
- Fraichard T (1993) Dynamic trajectory planning with dynamic constraints: a ‘state-time space’ approach. In: *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems*, Yokohama, doi:10.1109/IROS.1993.583794
- Fraichard T (2007) A short paper about motion safety. In: *IEEE international conference on robotics and automation*, Rome, doi:10.1109/ROBOT.2007.363138
- Fraichard T, Asama H (2004) Inevitable collision states. A step towards safer robots? *Adv Robot* 18(10):1001–1024
- Frazzoli E, Feron E, Dahleh M (2002) Real-time motion planning for agile autonomous vehicle. *AIAA J Guid Control Dyn* 25(1):116–129
- Haddad H, Khatib M, Lacroix S, Chatila R (1998) Reactive navigation in outdoor environments using potential fields. In: *Proceedings of the 1998 IEEE conference on robotics and automation*, Leuven, vol 2, pp 1232–1237
- Hart P, Nilsson N, Raphael B (1968) A formal basis for the heuristic determination of minimum-cost paths. *IEEE Trans Syst Sci Cybern* 4(2):100–107
- Howard T (2009) Adaptive model-predictive motion planning for navigation in complex environments. PhD thesis, Carnegie Mellon University
- Howard T, Kelly A (2007) Rough terrain trajectory generation for wheeled mobile robots. *Int J Robot Res* 26(2):141–166
- Howard T, Green C, Kelly A, Ferguson D (2008) State space sampling of feasible motions for high-performance mobile robot navigation in complex environments. *J Field Robot* 25(6–7):325–345
- Hsu D, Kindel R, Latombe JC, Rock S (2002) Randomized kinodynamic motion planning with moving obstacles. *Int J Robot Res* 21(3):233–255
- Iagnemma K, Shimoda S, Shiller Z (2008) Near-optimal navigation of high speed mobile robots on uneven terrain. In: *Proceedings of the 2008 international conference on robotics and automation*, Pasadena, 2008
- Jackson J, Crouch P (1991) Curved path approaches and dynamic interpolation. In: *IEEE aerospace and electronic systems magazine*, Glendale
- Jochem T, Pomerleau D, Kumar B, Armstrong J (1995) PANS: a portable navigation platform. In: *IEEE intelligent vehicles symposium*, Detroit, doi:10.1109/IVS.1995.528266
- Kalisiak M, van de Panne M (2007) Faster motion planning using learned local viability models. In: *IEEE international conference on robotics and automation*, Rome. doi:10.1109/ROBOT.2007.363873
- Kalmár-Nagy T, D’Andrea R, Ganguly P (2004) Near-optimal dynamic trajectory generation and control of a omni-directional vehicle. *Robot Auton Syst* 46:47–64
- Kanayama Y, Hartman B (1989) Smooth local path planning for autonomous vehicles. In: *Proceedings of the 1989 international conference on robotics and automation*, Santa Barbara, vol 3, pp 1265–1270
- Kanayama Y, Miyake N (1985) Trajectory generation for mobile robots. In: *Proceedings of the international symposium on robotics research*, Gouvieux, pp 16–23
- Kelly A, Nagy B (2003) Reactive nonholonomic trajectory generation via parametric optimal control. *Int J Robot Res* 22(7):583–601
- Kelly A, Stentz T (1998) Rough terrain autonomous mobility - part 2: an active vision and predictive control approach. *Auton Robot* 5:163–198

- Kelly A, Stentz T, Amidi O, Bode M, Bradley D, Mandelbaum R, Pilarski T, Rander P, Thayer S, Vallidis N, Warner R (2006) Toward reliable off-road autonomous vehicles operating in challenging environments. *Int J Robot Res* 25(5):449–483
- Khatib O (1986a) Real-time obstacle avoidance for manipulators and mobile robots. *Int J Robot Res* 5(1). doi:10.1177/027836498600500106
- Khatib O (1986b) Real-time obstacle avoidance for manipulators and mobile robots. *Int J Robot Res* 5(1):90–98
- Knepper R, Srinivasa S, Mason M (2010) An equivalent relation for local path sets. In: *Proceedings of the ninth international workshop on the algorithmic foundations of robotics*, Singapore
- Koenig S, Likhachev M (2002) D^{*} lite. In: *Proceedings of the AAAI conference on artificial intelligence*, Edmonton
- Komoriyama K, Taniguchi K (1989) Trajectory design and control of a wheel-type mobile robot using b-spline curve. In: *Proceedings of the 1989 IEEE/RSJ international conference on intelligent robots and systems*, Tsukuba, pp 398–405
- Kuwata Y, Karaman S, Teo J, Frazzoli E, How J, Fiore G (2009) Real-time motion planning with applications to autonomous urban driving. *IEEE Trans Contr Syst Technol* 17(5). doi:10.1109/TCST.2008.2012116
- Lavalle S (1998) Rapidly-exploring random trees: a new tool for path planning. Technical report, 98-11, Department of Computer Science, Iowa State University
- Lavalle S (2006) *Planning algorithms*. Cambridge University Press. <http://planning.cs.uiuc.edu/>. Accessed 26 Sep 2011
- LaValle S, Kuffner J (1999) Randomized kinodynamic planning. In: *IEEE international conference on robotics and automation*, Detroit, doi:10.1109/ROBOT.1999.770022
- Lozano-Perez T (1983) Spatial planning, a configuration space approach. *IEEE Trans Comput* 32(2):108–120
- Minguez J, Montano L (2004) Nearness diagram (ND) navigation: collision avoidance in troublesome scenarios. *IEEE Trans Robot Autom* 20(1):45–59
- Mitchell I, Tomlin C (2003) Overapproximating reachable sets by hamilton-jacobi projections. *J Sci Comput* 19(1–3). doi:10.1023/A:1025364227563
- Montemerlo M, Becker J, Bhat S, Dahlkamp H, Dolgov D, Ettinger S, Haehnel D, Hilden T, Hoffmann G, Huhnke B, Johnston D, Klumpp S, Langer D, Levandoski A, Levinson J, Marzil J, Orenstein D, Paefgen J, Penny I, Petrovskaya A, Pflueger M, Stanek G, Stavens D, Vogt A, Thrun S (2008) Junior: the stanford entry in the urban challenge. *J Field Robot* 25(9). doi:10.1002/rob.20258
- Murray R, Sastry S (1993) Nonholonomic motion planning: steering using sinusoids. *IEEE Trans Autom Contr* 38:700–716
- Petti S, Fraichard T (2005) Safe motion planning in dynamic environments. In: *Proceedings of the IEEE-RSJ international conference on intelligent robots and systems*, Edmonton
- Pivtoraiko M, Knepper R, Kelly A (2009) Differentially constrained mobile robot motion planning in state lattices. *J Field Robot* 26(3):308–333
- Prajna S, Jadbabaie A, Pappas G (2007) A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Trans Autom Contr* 52(8). doi:10.1109/TAC.2007.902736
- Reeds J, Shepp L (1990) Optimal paths for a car that goes both forwards and backwards. *Pacific J Math* 145(2):367–393
- Reichardt D, Shick J (1994) Collision avoidance in dynamic environments applied to autonomous vehicle guidance on the motorway. In: *IEEE intelligent vehicles symposium*, New York, doi:10.1109/IVS.1994.639475
- Reif J, Sharir M (1985) Motion planning in the presence of moving obstacles. In: *IEEE symposium on the foundations of computer science*, Cambridge, doi:10.1109/SFCS.1985.36
- Rogers-Marcovitz F, Kelly A (2010) On-line mobile robot model identification using integrated perturbative dynamics. In: *Proceedings of the 2010 international symposium on experimental robotics*, Delhi
- Rohrmuller F, Althoff M, Wollherr D, Buss M (2008) Probabilistic mapping of dynamic obstacles using markov chains for replanning in dynamic environments. In: *IEEE/RSJ international conference on intelligent robots and systems*, Nice, doi:10.1109/IROS.2008.4650952
- Schmidt C, Oechsle F, Branz W (2006) Research on trajectory planning in emergency situations with multiple objects. In: *IEEE intelligent transportation systems conference*, Toronto, doi:10.1109/ITSC.2006.1707153

- Seder M, Petrovic I (2007) Dynamic window based approach to mobile robot motion control in the presence of moving obstacles. In: IEEE international conference robotics and automation, Roma
- Shiller Z, Chen J (1990) Optimal motion planning of autonomous vehicles in three-dimensional terrains. In: Proceedings of the IEEE international conference on robotics and automation, Cincinnati, pp 198–203
- Shin D, Singh S (1991) Path generation for robot vehicles using composite clothoid segments. Technical report, Carnegie Mellon University
- Song D, Lee H, Yi J, Levandowski A (2007) Vision-based motion planning for an autonomous motorcycle on ill-structured roads. *Auton Robot* 23(3):197–212
- Stentz A, Herbert M (1995) A complete navigation system for goal acquisition in unknown environments. *Auton Robot* 2(2):127–145
- Stentz A, Kelly A, Herman H, Rander P, Amidi O, Mandelbaum R (2002) Integrated air/ground vehicle system for semi-autonomous off-road navigation. In: Proceedings of the AUVSI unmanned systems symposium, Orlando, 2002
- Thrun S (2010) What we're driving at. The official Google blog. <http://googleblog.blogspot.com/2010/10/what-were-driving-at.html>. Accessed 26 Sep 2011
- Thrun S, Montemerlo M, Dahlkamp H, Stavens D, Aron A, Diebel J, Fong P, Gale J, Halpenny M, Hoffman G, Lau K, Oakley C, Palatucci M, Pratt V, Stang P, Strohband S, Dupont C, Jendrossek L, Koelen C, Markey C, Rummel C, van Niekerk J, Jensen E, Alessandrini P, Bradski G, Davies B, Ettinger S, Kaehler A, Naflan A, Mahoney P (2006) Stanley: the robot that won the darpa grand challenge. *J Field Robot* 23(8):661–692
- Tilbury D, Laumond J, Murray R, Sastry S, Walsh G (1992) Steering car-like systems with trailers using sinusoids. In: Proceedings of the 1992 IEEE international conference on robotics and automation, Nice
- Ulmer B (1992) VITA-an autonomous road vehicle (ARV) for collision avoidance in traffic. In: IEEE intelligent vehicles symposium, New York, doi:10.1109/IVS.1992.252230
- Ulmer B (1994) VITA II-active collision avoidance in real traffic. In: IEEE intelligent vehicles symposium, New York, doi:10.1109/IVS.1994.639460
- Ulrich I, Borenstein J (2000) VFH*: local obstacle avoidance with look-ahead verification. In: IEEE international conference on robotics and automation, Washington, DC, doi:10.1109/ROBOT.2000.846405
- Urmson C, Ragusa C, Ray D, Anahlt J, Bartz D, Galatali T, Gutierrez A, Johnston J, Harbaugh S, Kato H, Messner W, Miller N, Peterson K, Smith B, Snider J, Spiker S, Ziglar J, Whittaker W, Clark M, Koon P, Mosher A, Struble J (2006) A robust approach to high-speed navigation for unrehearsed desert terrain. *J Field Robot* 23(8):467–508
- van den Berg J, Overmars M (2008) Planning time-minimal safe paths amidst unpredictably moving obstacles. *Int Journal of Robotics Research* 27(11–12). doi:10.1177/0278364908097581
- Vasquez D (2007) Incremental learning for motion prediction of pedestrians and vehicles. PhD thesis, Inst. Nat. Polytechnique de Grenoble. <http://tel.archives-ouvertes.fr/tel-00155274>. Accessed 26 Sep 2011
- Vasquez D, Fraichard T, Laugier C (2009) Growing hidden Markov models: a tool for incremental learning and prediction of motion. *Int J Robot Res* 28(11–12):1486–1506
- Vatcha R, Xiao L (2008) Perceived CT-space for motion planning in unknown and unpredictable environments. In: Workshop on algorithmic foundations of robotics, Guanajuato
- Wettergreen D, Tompkins P, Urmson C, Wagner M, Whittaker W (2005) Sun-synchronous robotics exploration: technical description and field experimentation. *Int J Robot Res* 24(1):3–30
- Xia T, Yang M, Yang R, Wang C (2010) Cyberc3: a prototype cybernetic transportation system for urban applications. *IEEE Trans Intell Transp Syst* 11(1). doi:10.1109/ITITS.2009.2036151

56 Risk Based Navigation Decisions



Anne Spalanzani¹ · Jorge Rios-Martinez² · Christian Laugier² · Sukhan Lee³

¹UPMF-Grenoble 2/INRIA Rhône-Alpes/Lig UMR, Grenoble, France

²e-Motion Project-Team, INRIA Rhône-Alpes, Saint Ismier Cedex, France

³School of Information and Communication Engineering, Department of Interaction Science, ISRI (Intelligent Systems Research Institute), Sungkyunkwan University, Jangan-gu, Suwon, Rep. of Korea (South Korea)

1	<i>Introduction</i>	1460
1.1	Problem Description	1460
1.2	Predicting the Future	1461
1.3	Example of Prediction with Gaussian Processes	1462
2	<i>Navigation Using Models of the Future</i>	1462
2.1	Related Work	1462
2.2	Focus on Risk-RRT Method	1464
2.2.1	The Risk-RRT Algorithm	1465
2.2.2	Risk Guided Search	1467
2.2.3	Real-Time Decisions Update	1468
3	<i>Human Aware Navigation</i>	1470
3.1	Related Work	1471
3.2	Adding Comfort Constraint to Risk Function	1471
3.2.1	Personal Space	1472
3.2.2	F-formations	1472
3.2.3	Model of O-space in F-formations	1473
3.2.4	Adding Social Constraints to Risk-RRT	1474
4	<i>Conclusions</i>	1475

Abstract: This chapter addresses autonomous navigation in populated and dynamic environments. Unlike static or controlled environments where global path planning approaches are suitable, dealing with highly dynamic and uncertain environments requires to address simultaneously many difficult issues: the detection and tracking of the moving obstacles, the prediction of the future state of the world, and the online motion planning and navigation. In the last few years, the problem of incomplete, uncertain, and changing information in the navigation problem domain has gained even more interest in the robotic community and probabilistic frameworks aiming to integrate and elaborate properly such information have been developed. This chapter is divided into three sections: First section introduces the main challenge of this approach.  [Section 2](#) focuses on navigation using prediction of the near future and  [Sect. 3](#) discusses on integrating human in the navigation decision scheme.

1 Introduction

Autonomous robots are widely spread in industries, where their space of work is well protected, so that nothing can interfere with the moving robot. In order to introduce robots in everyday life environments, safe systems of navigation in dynamic and populated environments have to be developed. In the last decade, a variety of mobile robots designed to operate autonomously in environments populated by humans has been developed. These robots have been deployed in hospitals, office buildings, department stores, and museums. Existing robotic systems are able to perform various services such as delivery, education, providing tele-presence, cleaning, or entertainment. Furthermore, there are prototypes of autonomous wheelchairs and intelligent service robots which are designed to assist people in their homes.

Autonomous navigation in populated environments still represents an important challenge for robotics research. In contrast with static or controlled environments where global path planning approaches are suitable, highly dynamic environments present many difficult issues: the detection and tracking of the moving obstacles, the prediction of the future state of the world, and the online motion planning and navigation. The decision about motion must be related with the online perception of the world, and take into account all the sources of uncertainty involved. In the last few years, the problem of incomplete, uncertain, and changing information in the navigation problem domain has gained more and more interest in the robotic community, and probabilistic frameworks aiming to integrate and elaborate such information have been developed.

1.1 Problem Description

The purpose is to develop techniques for a robot to move autonomously and safely in an environment which is not perfectly known *a priori* and in which static and moving obstacles are present. The task of the robot is to find and execute a sequence of actions

to reach a given position, avoiding collisions with the obstacles. The aim is to give a robot the possibility to exploit the fact that pedestrians and vehicles usually do not move at random in the given environment but often engage in typical behaviors or motion patterns. The robot may use this information to better predict the future position of these moving obstacles and adapt its behavior accordingly. The robot may also follow social convention to be well integrated in the human-populated environment. To develop methods for safe autonomous navigation among static and moving obstacles, some considerations must be taken:

- The fact that the environment is *dynamic* cannot be ignored: the robot performance is influenced by obstacles moving in the environment and the robot must be able to take safe and good decisions at anytime and act promptly in the dynamic environment.
- The *uncertainty* and *incompleteness* of the information perceived by the robot is not negligible and some mean to take it into account into the decision process should be introduced.

The main differences between methods presented here and classical planning methods are:

- Finding the shorter path is not the main objective.
- Navigation decisions are based on a risk evaluation. The risk function can rely on safety but also on comfort and human friendly navigation.

1.2 Predicting the Future

In a given environment, pedestrians and vehicles often engage in typical behaviors or motion patterns. Supposing that the environment has been observed for enough time and that the typical patterns have been learned, the information gathered provides a more reliable prediction in the medium and long-term with respect to a simple linear and conservative model (Vasquez Govea 2007), and a hint of the zones from where new obstacles are likely to enter the scene. The learning of typical patterns and the representation of pattern based motion models has been the subject of extensive study and many different approaches have emerged. Typical trajectories are usually represented as a sequence of points in the continuous state space. Most approaches do not model the time variable explicitly and assume that the points in the sequence are regularly spaced in time. Sometimes, a measure of the “width” of the cluster is also included in the representation (Makris and Ellis 2001; Junejo et al. 2004). In (Vasquez Govea 2007), a probabilistic model is proposed in which the width of the trajectory is represented as the variance of the distance between the trajectories that belong to the same cluster. Another probabilistic model of width has been proposed by (Bennewitz et al. 2005): every point of the trajectory prototype is modeled as a Gaussian and it is assumed that all such Gaussians have the same covariance matrix. A novel approach has been proposed by (Tay and Laugier 2007) where trajectories are represented by Gaussian Processes. In this case,

both the typical trajectory and its “width” (mean and covariance) are probabilistically estimated by a proper Bayesian framework. The advantages of the Gaussian Processes representation are that they present a solid probabilistic theory for the representation of the mean and covariance of the different paths and the future prediction. Also, trajectories are represented by continuous functions, which allow to use different time steps for prediction and for observation, thereby limiting the necessary interpolations at the learning phase. Finally, the Gaussian representation allows the prediction to be very fast and computationally cheap.

1.3 Example of Prediction with Gaussian Processes

► [Figure 56.1](#) shows the Gaussian mixture prediction at four different timesteps. Column (a) shows the environment, the observation points (red dots), and the prediction obtained. The gray lines show the means of all the typical patterns of the environment; the colored lines are the patterns that are retained for prediction after the gating. The ellipses represent the Gaussian mixture predicted for the next 10 s with a discretization of 0.5 s. The center of one ellipse is on the mean of the Gaussian component of the prediction and its radius is equivalent to one time its covariance. Column (b) shows the estimated likelihood for the Gaussian Processes retained for prediction. In the first timestep, all the trajectories originating at the door where the pedestrian is observed are likely. The prediction gets more precise as the history of observations gets longer: after some more observations, only the patterns going toward the right are retained and in the following timesteps, the red path becomes prominent with respect to the others. More details about this method can be found in chapter “Vehicle Prediction and Risk Assessment” and (Tay and Laugier [2007](#)).

2 Navigation Using Models of the Future

2.1 Related Work

Literature on navigation with pattern based motion models is quite poor. An early work on navigation in changing environments and in presence of typical motion suggests to divide the state space in *hazardous* and *shelter* regions (LaValle and Sharma [1997](#)). A shelter designates an area in which the robot is guaranteed to avoid collision, while a hazardous region designates an area in which other obstacles can move. The cost of traversing an hazardous or dynamic region directly corresponds to the risk of encountering a moving obstacle.

In more complex environments however, this representation may reveal too simplistic: there may be no shelter areas at all, or they can be interleaved with the hazardous ones, so that having a spatial and temporal hint of where moving object actually becomes a necessity.

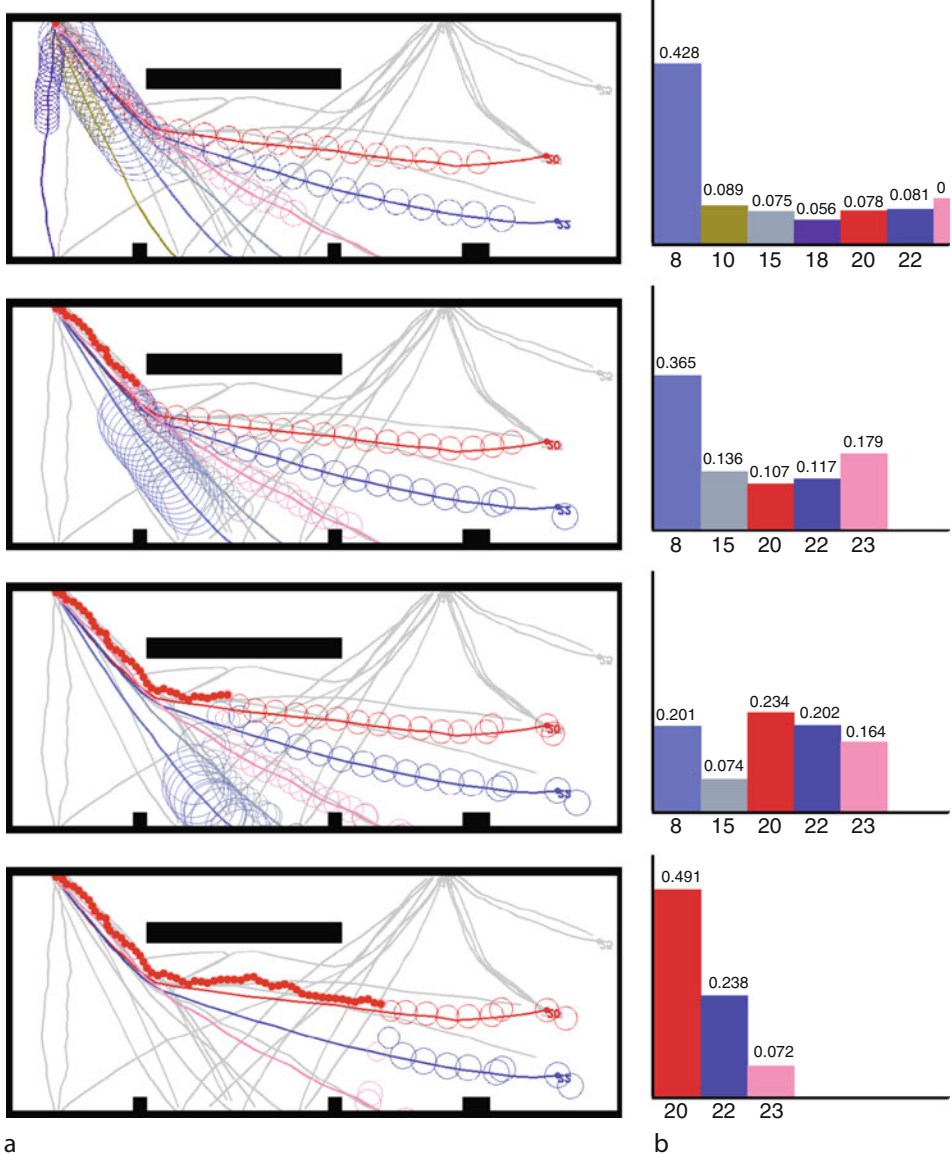



Fig. 56.1

Trajectory prediction based on Gaussian Processes representation: (a) shows the considered observation history (*solid dots*) and the prediction obtained for 15 time-steps ahead: *circles* are centered on the mean of each Gaussian component and have radius equal to one time the standard deviation. (b) shows the likelihood corresponding to each GP in the mixture. Only likelihoods bigger than 0.05 are shown

In (Bennewitz and Burgard 2003), the robot applies an A^* algorithm to find a path on a 2D space cost grid: the cost of passing through a cell at time t is given by the probability of collision plus the probability that a person covers the cell at that time. The algorithm is applied in an office-like environment where each typical pattern is represented by a fixed number of Gaussians that specify the probability of a position at a stage of the trajectory. Replanning is performed whenever information changes. In (Bennewitz et al. 2005), the use of these motion models is shown to improve the navigation performance in comparison with the case of a model based only on target tracking. In (Aoude et al. 2010), the authors propose to integrate a prediction of the intention of mobile obstacles (which are human driver vehicles) in the navigation decision. This Threat-Aware Path Planning method is an extension of the RRT algorithm (the CL-RRT). (Fulgenzi 2009) combines an RRT algorithm with models of prediction based either on Kalman Filter, GHMM (Fulgenzi et al. 2009), or Gaussian Processes (Fulgenzi et al. 2008). This method is described in  Sect. 2.2.

2.2 Focus on Risk-RRT Method

This method addresses particularly the problem of putting in relation the decision and action process with a realistic perception input, assuming that the robot has little *a priori* knowledge about the static environment and about the surrounding moving obstacles. The problems and limits inherent to sensor perception and future prediction are investigated but also environmental models that best express the changing information and uncertainty coming from perception. The unknown dynamic environment is perceived and mapped by the robot during navigation. To properly map the uncertainty and incompleteness of the sensing information regarding both static and dynamic environments probabilistic representations are chosen. The static environment is mapped by an occupancy grid (Elfes 1989); moving obstacles are detected and their dimension, position, and velocity are estimated thanks to a target tracking method. This method combines a prediction of the near future using typical patterns and a navigation algorithm based on risk assessment.

Risk-RRT (Fulgenzi 2009) is an autonomous navigation method based on the well-known RRT framework (LaValle and Kuffner 2007), designed to operate in dynamic, uncertain environment. Risk-RRT incorporates a probabilistic risk of collision as guide for searching safe paths which conduct the robot to its goal. The probabilistic risk of collision is computed on the basis of the probabilistic models which represent the static and dynamic obstacles. The search strategy is integrated in an anytime planning and replanning approach: the probabilities of collision and the decisions of the robot are updated online with the most recent observations. This latter feature permits the algorithm to plan taking into account real-time constraint. The navigation proposed supposes that dynamic obstacles, in a particular environment, move following typical patterns. These patterns can be learned (Vasquez Govea 2007; Meng Keat Christopher 2009) and used to better predict the future motion of dynamic obstacles.

At a given time, the robot knowledge about the state of the world is represented by:

1. An estimation of the state of the robot:

$$[x, y, \theta, v, \omega]^T$$

where (x, y, θ) is the position and orientation of the robot in the plane and (v, ω) are its linear and angular velocity respectively.

2. A set of Gaussian Processes which represent the typical patterns of the obstacles:

$$\mathcal{G} = \{G_k\}_{k=1 \dots K}$$

with K the total number of known typical patterns.

3. A goal position:

$$g = [x, y]^T$$

4. An occupancy grid which represents the structure of the static environment around the robot according to the previous observations:

$$\mathcal{M}(t) = \{p_{occ}(x, y)\}_{x \in X, y \in Y}$$

where p_{occ} is the probability of occupation, X and Y are finite sets representing the discrete coordinates of the cells of the grid.

5. A list of moving objects their estimated position, velocity, and previous observations:

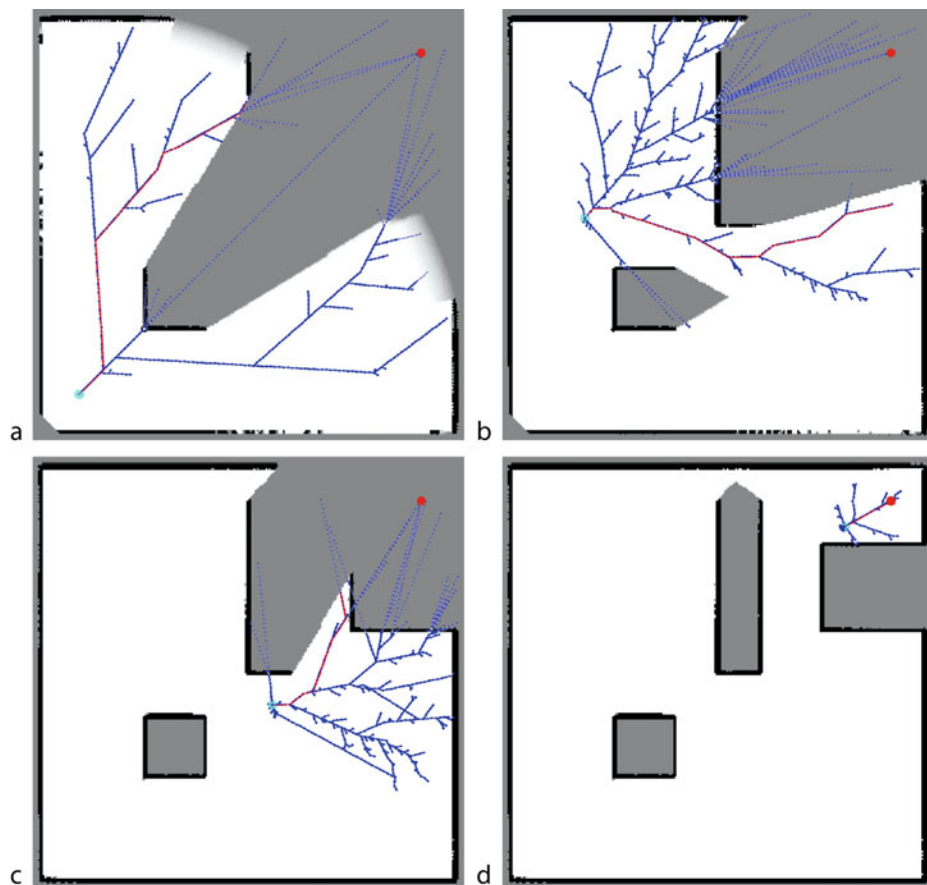
$$O = \{o_m\}_{m=1 \dots M} = \{[x, y, vx, vy]_m^T, H_m\}_{m=1 \dots M}$$

where H_m represents the history of observation related to obstacle o_m .

The occupancy grid and the state of the robot are estimated by a Simultaneous Localization and Mapping algorithm based on scan matching; the position, velocity, and track of the obstacles are estimated thanks to a Multiple Target Tracking algorithm (Trung-Dung and Olivier Aycard 2007). The typical patterns are supposed to have been learned by an off-board platform before navigation and to be known by the robot.

2.2.1 The Risk-RRT Algorithm

The motion planning algorithm proposed is described in [Algorithm 2.2.1](#). It combines a task dedicated to perception (of static and moving obstacles), a task for planning partial but safe trajectories and another one for navigating along planned safe trajectories. In practical, navigation and planning are done in parallel. The prediction done for forecasting the position of moving obstacles in the near future can be done by different ways, depending on the knowledge the robot has on the environment. If the robot does not have any model of the future, the behaviors of the mobile obstacle are considered as



■ Fig. 56.2

Risk-RRT: Growing tree evolving in time

conservative and a short-term prediction is used for the planning process. (Fulgenzi et al. 2009). If the robot has models of the future, a long-term prediction using these models is used. (Tay and Laugier 2007) describes Gaussian Processes based method for predicting moving obstacles trajectories. The selection of the best trajectories is done by computing a probability of collision between the robot following these trajectories and the moving obstacles following the predicted trajectories.

In ► Fig. 56.2, the tree of trajectories generated by the algorithm is shown at four instants during navigation. The initial position of the robot is at the left bottom corner, while the goal is at the right top corner. At the beginning, the most likely paths are explored in the two possible directions and the most promising one is chosen: the more promising path is drawn in red in ► Fig. 56.2a. ► Figure 56.2b shows the tree after some steps: the tree has been updated: the branch in the right direction has been cut as it became unreachable and the tree has been grown a little toward the promising direction.

► [Figure 56.2c](#) and [d](#) shows the tree and the new partial path found when a bigger portion of the space is visible.

Risk-RRT

```

1: procedure Risk-RRT
2:   trajectory = empty
3:   Tree = empty
4:   Goal = read()
5:   t = clock()
6:   while Goal not reached do
7:     if trajectory is empty then
8:       brake
9:     else
10:      move along trajectory for one step
11:    end if
12:    observe ( $X$ );
13:    delete unreachable trajectories ( $T, X$ )
14:    observe(Map movingobstacles)
15:    t = clock()
16:    predict moving obstacles at time  $t, \dots, t + N\tau$ 
17:    if environment different then
18:      update trajectories( $T$ , Map, moving obstacles)
19:    end if
20:    while clock() <  $t + \tau$  do
21:      grow trajectories with depth  $\leq N$  in  $T$ 
22:    end while
23:    trajectory = Choose best trajectory in  $T$ 
24:    t = clock()
25:  end while
26:  brake
27: end procedure

```

2.2.2 Risk Guided Search

This paragraph explains how the configuration-time space is searched and how a path is chosen. These operations correspond respectively to line 21 and line 23 in ► [Algorithm 2.2.1](#).

The configuration-time space is searched randomly and a tree T is grown from the initial configuration all over the configuration space. The algorithm chooses a point P in the configuration space and tries to extend the current search tree toward that point. In the classical RRT algorithm, P is chosen randomly in the free configuration space. In this problem, there is little or no knowledge on the structure of the environment: P is sampled

from the rectangular region between the robot and the goal enlarged by some amount to take into account for possible local minima. P can be in an occupied or in an unknown zone. At the beginning, and then once on 100 times, the goal is chosen; this bias, which has been empirically set, speeds up the exploration toward the goal. The node chosen for extension is the most *promising* node: all the nodes in T are weighted taking into account the risk of collision and the estimated length of the total path:

$$\tilde{w}(q_N) = \frac{L_\pi(q_N)}{\text{dist}(q_0, q_N, P)} \quad (56.1)$$

$$w(q_N) = \frac{\tilde{w}(q_N)}{\sum_q \tilde{w}_q} \quad (56.2)$$

At numerator, the likelihood of $\pi(q_N)$, path from position q_0 to position q_N , is normalized with respect to the length of the path N ; at denominator, $\text{dist}(\cdot)$ is the sum between the length of the path from the root q_0 to the node q_N (which is known) and an estimation of the length of the path to P :

The weights are normalized over the set of nodes in the tree (► Eq. 56.2). The node to grow next is then chosen taking the maximum over the weights or drawing a random node proportionally to the weight. The new node q_+ is obtained applying an admissible control from the chosen node q toward P . The weight of q_+ is computed. If $w(q_+) \geq w(q)$ the tree is grown again from q_+ toward P otherwise another point is sampled from the space. When the available time for planning is over, the best partial path with the highest weight for the goal is retrieved and passed to execution.

2.2.3 Real-Time Decisions Update

This paragraph explains how the search tree is updated and how the information coming from perception is integrated in it (see respectively lines 13 and 18 of ► Algorithm 2.2.1).

In a dynamic environment the robot has a limited time to perform planning which depends on the time-validity of the models used and on the moving objects in the environment. The conditions used for planning could be invalidated at execution time: for example an obstacle could have changed its behavior or some new obstacle could have entered the scene. The idea of Partial Motion Planning (Petti and Fraichard 2005) is to take explicitly into account the real-time constraint and to limit the time available for planning to a fixed interval. After each planning cycle, the planned trajectory is generally just a partial trajectory. Execution and planning are done in parallel: while the robot moves a step along the planned partial path, the tree is updated with the information coming from the perception algorithm, the tree is grown and the new partial path is passed for execution when the timestep is over. In order to do this, the expected state of the robot at next step becomes the root of the new search tree. If there is no error in the execution of the

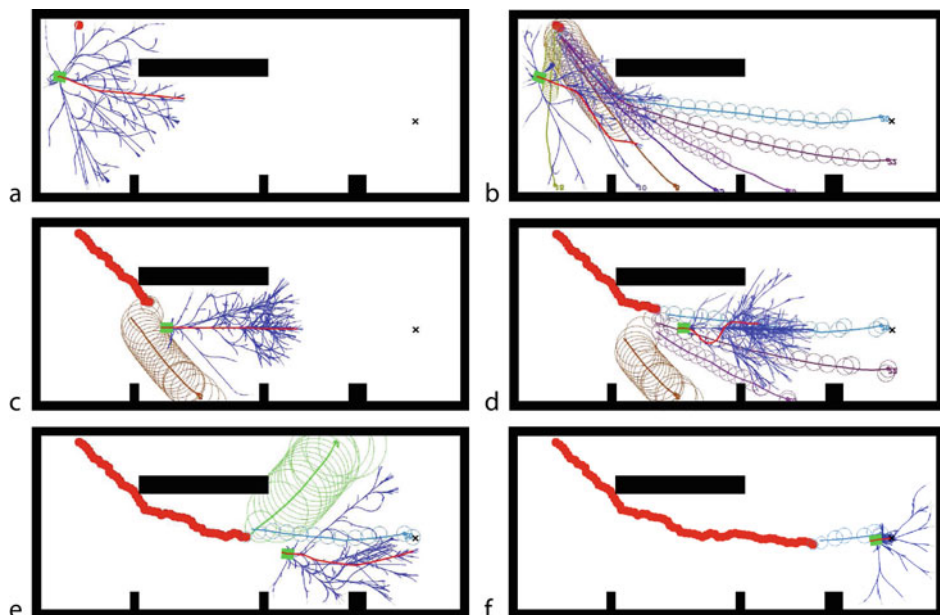
robot, the robot expected state is a node along the best partial path already explored. In this case, the subtree of this node becomes the current search tree. If there is some error and the robot is far away of the configuration it is expected to reach, the new tree is constituted only by the new expected position and search must begin from scratch. The updating step is done in two phases. The likelihood of each partial path can also be expressed as the multiplication of the independent static and dynamic components:

$$L_{\pi}(q_N) = \prod_{n=0}^N (1 - P_{cs}(q_n)) \cdot \prod_{n=0}^N (1 - P_{cd}(q_n)) \quad (56.3)$$

this two values are both stored in the nodes, so that they can be updated separately and only when needed. When an observation comes in from the perception algorithm, the planner checks for differences in the observed grid at first and in the tracked obstacles then. The incoming grid and the old one are subtracted; where a difference is found, the P_{cs} of the corresponding node is updated. For the moving obstacles, the algorithm checks for difference in the weights w_m , k , and the P_{cd} of the affected nodes is correspondingly updated. The tree is grown in the rest of available time.

When the static environment is known and free of moving obstacles, the algorithm degenerates to a classical RRT approach: the likelihood of the nodes is either 0, when the corresponding space is occupied, either it depends only on the distance to the goal, which is the case for the deterministic RRT algorithm. Also, the new observations do not give any new information, so there is no need to update the tree and the search can continue till the goal is reached.

Figure 56.3 shows some snapshots from the obtained results. The robot is the green rectangle and perceives the circular obstacle (red full point). The goal of the robot is the black cross on the right. Colored lines represent the portion of the GPs means retained for prediction at each time from the nearest point to the last observation to the end. Colored ellipses represent the prediction, as explained in Figure 56.5. The tree explored by the robot is drawn by the blue lines. Lighter blue means lower likelihood. The red line is the path chosen each time. Figure 56.3a shows the planning at the first timestep, when no obstacle has been detected yet: the path leads straight toward the goal. In Figure 56.3b, an obstacle has been detected for the first time. Many patterns are possible: the probabilities of collision on the search tree are modified and the most likely path now gets the robot further from the obstacle. In Figure 56.3c, the pedestrian is believed to go toward the bottom part of the environment: the robot plans to drive directly toward the goal. Few timesteps after (Figure 56.3d) the prediction is modified: the pedestrian will go toward the right. The robot finds itself on the trajectory of the pedestrian: the trajectories searched are not safe. After some timesteps, a new solution is found: in Figure 56.3e, the robot plan is to get away from the path of the pedestrian before going toward the goal. In Figure 56.3f, the robot approaches the goal from upward to avoid crossing the trajectory of the obstacle. More details can be found in (Fulgenzi 2009).



■ Fig. 56.3

The robot moves in a simulated environment with a moving obstacle. The prediction of the obstacle is given by a Gaussian mixture based on the pre-learned Gaussian processes (*large gray squares*). The exploration tree maintains an estimation of the likelihood of the path that is adapted to the incoming observation

3 Human Aware Navigation

Robots enter more and more into human environments, humans will share the physical space with robots. Therefore robots need to take into account the presence of humans. Their trajectories must be safe but also predictable. Their behavior should follow social conventions, respecting proximity constraints, avoiding people interacting, or joining a group engaged in conversation without disturbing. The risk function can rely on safety but also in human friendly navigation. Once that collision can be successfully avoided, one next step follows in priority, insure the comfort of both users and people in the environment extending the problem of navigation to a human aware navigation.

People maintain a set of social conventions related to space when they are interacting, for example, in a conversation (Ciolek and Kendon 1980). The sociology literature often refers to the concept of personal space proposed by (Hall 1966) which characterizes the space around a human being in terms of comfort to social activity. Concerning interactions between people, the concept of o-space is described in the sociology literature. This space can be observed in casual conversations among people (Ciolek and Kendon 1980). Perception of territorial boundaries established by a group of humans and respect to these bounds are evidence of social behavior. If the robot aims to join a group, it must get permission from the group to be integrated.

In this section, a way to formulate comfort constraints is proposed. The concepts of personal space and o-space are included in the risk function proposed in the Risk-RRT approach. First, a review of some recent works which incorporate human aware navigation is presented.

3.1 Related Work

The literature shows the growing interest of the robotics community in research including proxemics and its impact in the development of tasks by the robot. In (Gockley et al. 2007), it is argued that moving in easily understood and predictable ways will both improve people's trusting and comfort with the robot as well as will help to insure the safety of people moving near the robot. They proposed a model for person following behavior and evaluated two approaches: one following the exact path of the person and other following in the direction of the person, they conclude that the second one is the most human-like behavior. The factors considered for the design of the model proposed are human-likeness, personal space, reliability in tracking of person, and safety.

In (Sisbot et al. 2007) a motion planner is presented which takes explicitly into account its human partners. The authors introduced criteria based both on the control of the distance between the robot and the human, and on the control of the position of the robot within the human's field of view.

In (Hansen et al. 2009) an adaptive system based on the person's pose and position, was introduced. This work, presented as a basis for human aware navigation, detects if a person seeks to interact with the robot. Navigation was implemented using human centered potential fields.

In (Kirby et al. 2009) a generalized framework for representing social conventions as components of a constraint optimization problem was presented and it was used for path planning and navigation. Social conventions were modeled as costs to the A^* planner with constraints like shortest distance, personal space, and pass on the right. Navigation was based in the Pure Pursuit Path-following. Simulation results showed the robot navigating in a "social" manner.

The work presented in (Chung and Huang 2010) proposed Spatial Behavior Cognition Model (SBCM), a framework to describe the spatial effects existing between human-human and human-environment. SBCM was used to learn and predict behaviors of pedestrians in an environment and to help a service robot to take navigation decisions. The algorithm Dynamic AO^* was used for motion planning issues.

3.2 Adding Comfort Constraint to Risk Function

In this section a way to add comfort constraints to risk-based navigation framework is proposed. The concepts of personal space and o-space as well as their relation with comfort are introduced and an extension to Risk-RRT is discussed.


3.2.1 Personal Space

The term Proxemics was proposed by (Hall 1966) to describe the use of space between humans, he observed the existence of some rules not written that conducted people to keep distances from others, and others to respect this space, he proposed that space around a person in social interaction is classified as follows:

- The public zone > 3.6 m
- The social zone > 1.2 m
- The personal zone > 0.45 m
- The intimate zone < 0.45 m

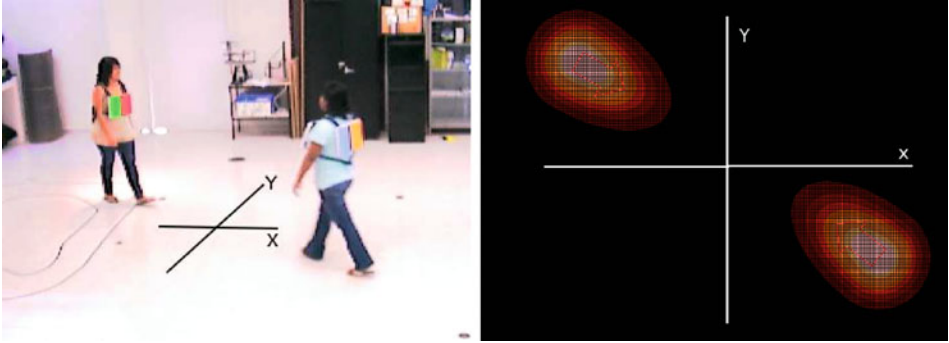
This definition is important because it represents a useful tool for a robot to understand humans intentions. It's well-known that these measures are not strict and that they change depending on age, culture, and type of relationship but the categories proposed explain very well reactions like the uncomfortable sense of a stranger invading your intimate zone or the perception of somebody looking social interaction by entering to your social zone. In general, people are more strict regarding their frontal space.

In the rest of the article, the word *personal space* is used as a synonymous of personal zone plus intimate zone.

The model implemented to represent personal space is defined in (Laga and Amaoka 2009), it consists in blending two Gaussian functions, both of them centered in the position of the person. The first one represents the personal space situated in front of a human and for this reason it's wider than the last one representing the back space.  **Figure 56.4** shows an example of personal space for two people walking, the measures are projected in the plane of floor, the values obtained from the Gaussian are higher in the center than on the borders.

3.2.2 F-formations

In 2010, Kendon people interact in groups and follow some spatial patterns of arrangement. When people are performing an activity they claim an amount of space related to that activity, this space is respected by other people and Kendon referred it as individual's *transactional segment*. This transactional segment varies depending on body size, posture, position, and orientation during the activity. Moreover, the groups can establish a joint or shared transactional segment and only participants have permitted access to it, they protect it and others tend to respect it. The *o-space* is that shared transactional segment reserved for the main activity. This space is surrounded by a narrower one, called the *p-space*, which provides for the placement of the participant's bodies and also personal things. An *F-formation* system is the spatial-orientation arrangement that people create, share and maintain around their o-space.



■ Fig. 56.4

Estimated personal space for two people that walk projected in the floor

3.2.3 Model of O-space in F-formations

As there is not an exact physical definition of o-space, this section describes how its location can be estimated.

When more than two people are in conversation they exhibit an F-formation with circular shape then the o-space could be taken as a circle whose center coincides with that of the inner space. In the case of two people, some F-formations have been identified as the most frequent (Ciolek and Kendon 1980) examples are shown in ● Fig. 56.5.

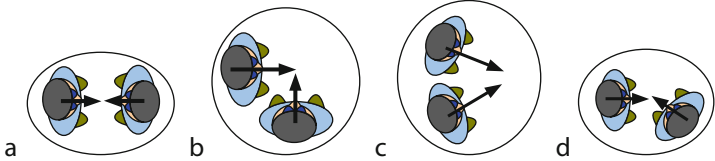
In this model, the o-space will be dependent on the particular F-formation identified: vis-a-vis, L-Shape, C-Shape, or V-Shape. From the definition found in the reference, a geometric representation for each F-formation can be designed, the model is based on the position and orientation of the body of the participants. Given the positions of pedestrians $H_1 = (x_1, y_1)$ and $H_2 = (x_2, y_2)$ in the plane of the floor and their respective orientations ϕ_1 and ϕ_2 around the normal to that plane, D_H is computed as the euclidean distance between H_1 and H_2 . A point V_i is also computed as the intersection of the vectors beginning in H_1 and H_2 in the direction of ϕ_1 and ϕ_2 , respectively. Let H_{12} be the mean point between H_1 and H_2 . Let C be the mean point between V_i and H_{12} . Calculate D_i as the distance between V_i and H_{12} .

The o-space is represented by a two-dimensional Gaussian function Γ_c of covariance matrix S and centered in C , then for each point Q around the center:

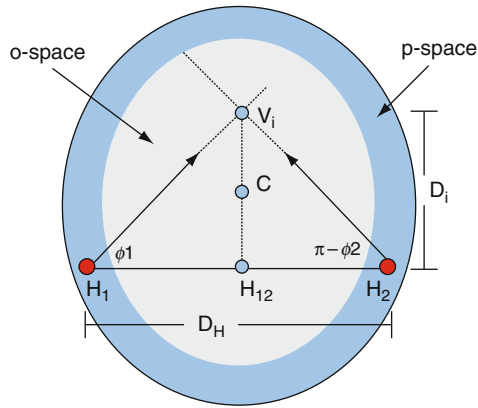
$$\Gamma_{C,S}(Q) = e^{-\frac{1}{2}(Q-C)^t S^{-1}(Q-C)} \quad (56.4)$$

where S is a diagonal covariance matrix defined as:

$$S = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}. \quad (56.5)$$



■ Fig. 56.5
Examples of F-formations: (a) Vis-a-vis, (b) L-shape, (c) C-shape, (d) V-shape



■ Fig. 56.6
Scheme showing the elements of the model O-space for L-shape F-formation

To get the shape of the o-space according to the F-formations, the values chosen for the parameters are $\sigma_x = D_H/4$ and $\sigma_y = D_i/2$. In the particular case of the Vis-a-Vis formation $\sigma_y = 0.6$. The orientation of the Gaussian is in the direction of the segment $\overrightarrow{H_{12}C}$, this coincides with the location of the point of interest of humans as exhibited by the orientation of their bodies.

All the elements defined can be seen in [Fig. 56.6](#) for the case of an L-Shaped F-formation.

3.2.4 Adding Social Constraints to Risk-RRT

Let's PZi defined as the probability of disturbing by passing inside the o-space ([Sect. 3.2.2](#)) of interaction *i*, it is computed as:

$$PZ_i = \max_{\varsigma} (\Gamma_{C_i, S_i}(Cell_{x,y})) \tag{56.6}$$

where ζ is the subset of cells which is the minimal approximation of surface A , the area swept by the robot.

Disturbing an interaction can be integrated in the risk computation as a collision with a dynamic obstacle. The probability of dynamic collision is then:

$$P_{cd} = 1 - \prod_{m=1}^M [1 - P_{cd}(o_m)] \prod_{i=1}^r [1 - PZ_i] \quad (56.7)$$

The probability P_{ps} of disturbing by passing in the personal space of the human o_m can be approximated by a probability that A intercepts the one represented by the personal space:

$$P_{ps}(o_m, k) = \int_A PS(o_m(t)) \quad (56.8)$$

Where $PS(o_m(t))$ is the model of personal space centered in $o_m(t)$ as described in [Sect. 3.2.1](#). Again, to take into account this last constraint, we need to modify the equation that calculates probability of collision with object o_m to get:

$$P_{cd}(o_m) = \sum_{k=1}^K w_{mk} [P_{cd}(o_m, k) + P_{ps}(o_m, k)(1 - P_{cd}(o_m, k))] \quad (56.9)$$

After these extensions the “probability of success” calculated for every partial path is given by the probability of not encountering a collision along the path and not entering in a personal space or an o-space. For more details about this method, refer to (Rios-Martinez et al. 2011).

4 Conclusions

This chapter proposed to integrate prediction of the future and social conventions to adapt the navigation strategy to the real world. It highlights the possibility and importance to take into account the knowledge about the behavior of moving obstacles such as pedestrians. Predicting the obstacles position on the basis of the typical patterns enables the robot to rely on a good knowledge of the environment configuration in the near future. The uncertainty of the prediction of the position of the objects has a finite dispersion and the robot has a more precise idea of where new obstacles may come from. From the point of view of navigation, this means that the robot can plan longer paths, the necessity of replanning is reduced and better global performance are observed (shorter global path, shorter time to reach the goal). Navigation taking into account social conventions were presented also, providing the robot the ability to respect social conventions followed by humans.

References

- Aoude GS, Luders BD, Levine DS, How JP (2010) Threat-aware path planning in uncertain urban environments. In: Proceedings of the 2010 IEEE/RSJ international conference on intelligent robots and systems, Taipei, 2010
- Bennewitz M, Burgard W (2003) Adapting navigation strategies using motion patterns of people. In: Proceedings of the IEEE international conference on robotics and automation, Taipei, pp 2000–2005. IEEE
- Bennewitz M, Burgard W, Cielniak G, Thrun S (2005) Learning motion patterns of people for compliant robot motion. *Int J Robot Res* 24(1): 31–48
- Chung SY, Huang HP (2010) A mobile robot that understands pedestrian spatial behaviors. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, pp 5861–5866
- Ciolek M, Kendon A (1980) Environment and the spatial arrangement of conversational encounters. *Sociol Inq* 50:237–271
- Elfes A (1989) Using occupancy grids for mobile robot perception and navigation. *Computer*, vol 22, pp 4657
- Fulgenzi C (2009) Autonomous navigation in dynamic uncertain environment using probabilistic models of perception and collision risk prediction. PhD thesis, Institut National Polytechnique de Grenoble – INPG. <http://tel.archives-ouvertes.fr/tel-00398055/en/>
- Fulgenzi C, Tay C, Spalanzani A, Laugier C (2008) Probabilistic navigation in dynamic environment using rapidly-exploring random trees and gaussian processes. In: IEEE/RSJ 2008 international conference on intelligent robots and systems, France Nice, 2008. <http://hal.inria.fr/inria-00332595/en/>
- Fulgenzi C, Spalanzani A, Laugier C (2009) Probabilistic motion planning among moving obstacles following typical motion patterns. In: IEEE/RSJ international conference on intelligent robots and systems, St. Louis, Missouri États-Unis d'Amérique, 2009. <http://hal.inria.fr/inria-00398059/en/>
- Gockley R, Forlizzi J, Simmons R (2007) Natural person following behavior for social robots. *Proceeding of HRI07*, 2007
- Hall ET (1966) The hidden dimension: man's use of space in public and private. The Bodley Head, London
- Hansen ST, Svenstrup M, Andersen HJ, Bak T (2009) Adaptive human aware navigation based on motion pattern analysis. In: The 18th IEEE international symposium on robot and human interactive communication, Toyama, 2009
- Junejo I, Javed O, Shah M (2004) Multi feature path modeling for video surveillance. In: Proceedings of the 17th international conference on pattern recognition, Cambridge, UK. August 2004. vol 2. pp 716–719
- Kendon A (2010) Spacing and orientation in co-present interaction. In: Development of multimodal interfaces: active listening and synchrony. *Lecture notes in computer science*, vol 5967. Springer, Berlin/Heidelberg, pp 1–15
- Kirby R, Simmons R, Forlizzi J (2009) Companion: a constraint-optimizing method for person acceptable navigation. In: The 18th IEEE international symposium on robot and human interactive communication, Toyama, Japan, 2009
- Laga H, Amaoka T (2009) Modeling the spatial behavior of virtual agents in groups for non-verbal communication in virtual worlds. In: IUCS '09, Tokyo, Japan, 2009
- LaValle S, Kuffner JJ Jr (2007) Randomized kinodynamic planning. *Int J Robot Res* 26: 997–1024
- LaValle SM, Sharma R (1997) On motion planning in changing, partially-predictable environments. *Int J Robot Res* 16:775–805
- Makris D, Ellis T (2001) Finding paths in video sequences abstract. In: British machine vision conference, Manchester, 2001, pp 263–272
- Meng Keat Christopher T (2009) Analysis of dynamic scenes: application to driving assistance. PhD in computer science, Institut National Polytechnique de Grenoble (INPG)
- Petti S, Fraichard T (2005) Safe motion planning in dynamic environments. In: Proceedings of the 2005 IEEE/RSJ international conference on intelligent robots and systems, Alberta, Canada, 2005
- Rios-Martinez J, Spalanzani A, Laugier C (2011) Probabilistic autonomous navigation using rrt approach and models of human interaction. In: Proceedings of the 2011 IEEE/RSJ

- international conference on intelligent robots and systems, San Francisco, USA, 2011
- Sisbot EA, Marin-Urias LE, Alami R, Simeon T (2007) A human aware mobile robot motion planner. *IEEE Trans Robot* 23(5):874–883
- Tay C, Laugier C (2007) Modelling paths using gaussian processes. In: *Proceedings of the international conference on field and service robotics, Chamonix, 2007*. <http://emotion.inrialpes.fr/bibemotion/2007/TL07>
- Trung-Dung Vu, Olivier Aycard NA (2007) Online localization and mapping with moving object tracking in dynamic outdoor environments. In: *IEEE intelligent vehicles symposium, Istanbul, 2007*
- Vasquez Govea DA (2007) *Incremental Learning for Motion Prediction of Pedestrians and Vehicles*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble (Fr)

57 Probabilistic Vehicle Motion Modeling and Risk Estimation

Christopher Tay¹ · Kamel Mekhnacha¹ · Christian Laugier²

¹Probayes SAS, Inovalée Saint Ismier Cedex, France

²e-Motion Project-Team, INRIA Grenoble Rhône-Alpes, Saint Ismier Cedex, France

1	<i>Introduction</i>	1481
1.1	Toward Better Collision Warning	1482
1.2	Intuition	1484
1.3	Organization	1485
2	<i>Collision Risk Estimation</i>	1485
2.1	Overall Architecture	1485
2.2	Behavior Recognition and Modeling	1487
2.3	Realizations of Behaviors	1489
2.3.1	Gaussian Process Deformation Model	1490
2.3.2	Property	1492
2.3.3	Implementation Issues	1492
2.3.4	Chosen Implementation	1492
2.4	Predicting Vehicle Motion	1494
2.4.1	Conformal Transformation Between World Space and Canonical Space	1495
2.4.2	Inferring Probability Distribution on Future Motion	1496
2.4.3	Mapping Back to Real-World Coordinates	1497
2.5	Evaluation of Risk	1498
2.5.1	Risk of Trajectory Considering Behavior of One Vehicle Only	1499
2.5.2	Risk of trajectory against one vehicle with behaviors aggregated	1500
2.5.3	Aggregating Risk with Respect to All Vehicles	1501
2.5.4	Risk Associated with Driving Behavior	1502
3	<i>Experiments</i>	1503
3.1	Monte Carlo Simulation Validation	1503
3.1.1	Experimental Setup	1504
3.1.2	Results	1505

3.2 Driving Simulation 1508

3.2.1 Experimental Setup 1509

3.2.2 Results 1509

4 Conclusion 1515

Abstract: The development of autonomous vehicles garnered an increasing amount of attention in recent years. The interest for automotive industries is to produce safer and more user-friendly cars. A common reason behind most traffic accidents is the failure on the part of the driver to adequately monitor the vehicle's surroundings. This chapter addresses the problem of estimating the collision risk for a vehicle for the next few seconds in urban traffic conditions.

Current commercially available crash warning systems are usually equipped with radar-based sensors on the front, rear, or sides to measure the velocity and distance to obstacles. The algorithms for determining the risk of collision are based on variants of time-to-collision (TTC). However, it might be misleading in situations where the roads are curved and the assumption that motion is linear does not hold. In these situations, the risk tends to be underestimated. Furthermore, instances of roads which are not straight can be commonly found in urban environments, like the roundabout or cross-junctions.

An argument of this chapter is that simply knowing that there is an object at a certain location at a specific instance in time does not provide sufficient information to assess its safety. A framework for understanding behaviors of vehicle motion is indispensable. In addition, environmental constraints should be taken into account especially for urban traffic environments.

This chapter proposes a complete probabilistic model motion at the trajectory level based on the Gaussian Process (GP). Its advantage over current methods is that it is able to express future motion independently of state space discretization. Driving behaviors are modeled with a variant of the Hidden Markov Model. The combination of these two models provides a complete probabilistic model for vehicle evolution in time. Additionally a general method of probabilistically evaluating collision risk is presented, where different forms of risk values with different semantics can be obtained, depending on its applications.

1 Introduction

The main problem of this chapter concerns the estimation of the risk of collision of a vehicle. From the driver's point of view, the driver can obtain a general indication of the risk of collision for the next few seconds, warning the driver of unnoticed risks. The estimated risk of collision can also be used to aid an autonomous vehicle in choosing a suitable trajectory to minimize its risks.

For a completely autonomous vehicle, or even for a crash warning system, estimation of the risk of collision is a component of the complete system. The estimation of the risk of collision receives a set of processed sensor information from other modules of the complete system and it outputs risk values, which is to be interpreted by the application in context.

Throughout this chapter, the following set of processed sensor inputs are assumed to be available:

1. *Road geometry:* In order for the risk estimation to be aware of the road constraints, it must have geometrical information such as the width of the road and its curvature.

Such information can be obtained from specific algorithms which process raw information from camera images or lidars. Alternatively, it is also possible to obtain road geometry information given a Geographic Information System (GIS) with a pre-built map and a localization device such as the GPS.

2. *Target tracking*: The estimation of collision risk necessitates the detection and tracking of moving obstacles. The position and velocity of the moving obstacles can then be obtained.
3. *Detailed specified sensors*: Some examples of additional information which are not crucial but desirable are information such as the detection of the status of the signal lights of other vehicles as it is a strong indicator of the intention of other moving vehicles. It is also possible to have additional “virtual” sensors coming from further processed raw sensory data. An example would be the distance of the vehicle to the left or right lane border, which might indicate intentions to perform a lane change. Such information is highly informative in driving behavior recognition.

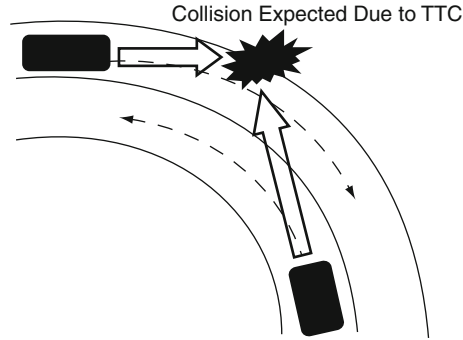
A vehicle which is referred to as the ego-vehicle is assumed to be equipped with the appropriate sensors so as to obtain the set of processed sensor inputs mentioned above. In this chapter, the estimated risk is a numerical value which expresses quantitatively the risk of the ego-vehicle going into collision with another vehicle in the next few seconds.

Estimating the risk of collision in the future involves the construction of models describing vehicle motion in the sensor visibility range of the ego-vehicle. Furthermore, this model should be capable of reasonably predicting future vehicle states. It is only with a prediction on future vehicle states that it is possible to estimate the risk of collision in the future.

When reasoning about the future, it is sensible to describe the future in terms of probability. A fully probabilistic vehicle evolution model for obtaining and inferring beliefs on the future states of vehicles in urban traffic environments is presented. Consequently, the estimated risk of collision can be obtained from the models in terms of probability in a theoretically consistent manner.

1.1 Toward Better Collision Warning

Current commercially available crash warning systems are mostly aimed at preventing front, rear, or side collisions. Such systems are usually equipped with radar-based sensors on the front, rear, or sides to measure the velocity and distance to obstacles. The algorithms for determining the risk of collision are based on variants of time-to-collision (TTC) (Lee 1976). TTC is basically a function of two objects, giving the time remaining before an object enters into collision with the other assuming that the two objects maintain the same linear velocity. Some systems are not passive but rather, it intervenes by directly controlling the brakes and possibly the steering to effectuate the



■ Fig. 57.1

Example of false collision alarm due to linear assumption of TTC based systems. Actual path of vehicles as dotted arrows. Source: (TAY 2009)

necessary corrective actions. Systems based on TTC are based on the fact that observations are made at a reasonably high frequency in order to adapt to potentially changing environments.

Current commercial systems work reasonably well on automotive highways or certain sections of the city where roads are straight. However, it might be misleading in situations where the roads are curved and thus, the assumption that motion is linear does not hold (see ► Fig. 57.1). In these situations, the risk tends to be underestimated. Furthermore, instances of roads which are not straight can be commonly found in urban environments, like the roundabout or cross-junctions.

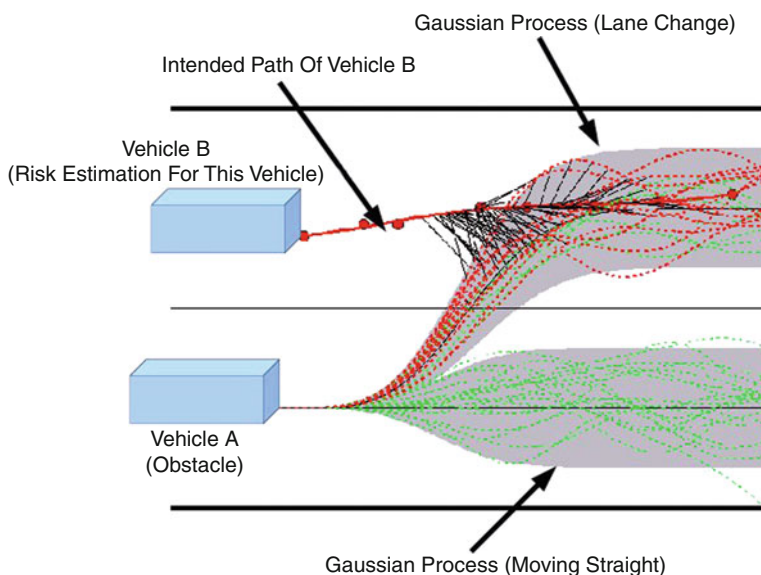
Several research projects were created to overcome such problems by taking into account the structure of the environment especially in intersections where there is a higher rate of accidents. These projects aim to provide intersection collision warning systems where there wireless communications are either between vehicles, or by using road side infrastructure such as traffic light information (Pierowicz et al. 2000; National Highway Traffic Safety 2004; Fuerstenberg and Chen 2007). Each of these systems has vehicles equipped with a pair of detectors (either radar sensors or laser scanners) at the left and right front corners of the vehicles in order to detect cross-traffic vehicles at intersections. The speeds and hence TTC of the obstacles are then evaluated to determine the risk of collision. Although the environmental structures are taken into consideration when evaluating the risk of collision, the actual calculation of the risk of collision is still based on the assumption of linear motion. The time horizon of risk prediction is short and crucial environmental information and information on sensors are not fully utilized.

An argument of this chapter is that simply knowing that there is an object at a certain location at a specific instance in time does not provide sufficient information to assess its safety. A framework for understanding behaviors of vehicle motion is indispensable. In addition, environmental constraints should be taken into account especially for urban traffic environments. An overview of the proposed approach is presented in the next section.

1.2 Intuition

An example of a possible scenario with two lanes both in the same direction as illustrated in [Fig. 57.2](#). Two vehicles A and B are traveling on separate lanes and the risk of collision is to be estimated for vehicle B. From the driver of vehicle A's point of view, the local road structure is implicitly described by maneuvers such as going straight, turning right/left, lane change. Such maneuvers shall be referred to as behaviors. The total set of possible behaviors are pre-defined. However, not all behaviors are available at all instances. For example, it might not be possible to turn left at the next intersection because there is no road turning left. The set of feasible behaviors at each instance is a subset of all the possible behaviors.

For each feasible behavior, there are a number of different ways of physically executing the behavior. Humans do not drive in an absolutely straight manner, precisely following the middle of the lane. However, it is reasonable to assume that a normal driving routine approximately follows the lane. The lane following for a given behavior is represented using a GP which gives a probability distribution over the possible future physical realizations of the paths where the mean will be the path following exactly the middle of the lane. The GPs and its variances are shown as gray regions in [Fig. 57.2](#) corresponding to behaviors lane changing and going straight. Dotted lines represent the path sampled



■ Fig. 57.2

Risk to be estimated of a trajectory to be taken by vehicle B. Path prediction for vehicle A (obstacle) is obtained by sampling from the GPs (one each for going straight and lane changing). The risk of collision is calculated by a weighted sum of trajectories in collision. Source: (TAY 2009)

from the GP. For cases where the road has a nonzero curvature or for turning behaviors, the GP will be appropriately adapted according to the geometry of the road.

The set of GPs for each of the feasible behavior in the scene, in combination with the probability that vehicle A executes a certain behavior, gives a probabilistic model of the future evolution of vehicle A in the scene.

Similar to the TTC evaluation of risk of collision, the evaluation of the risk of collision will be for vehicle B against vehicle A. In contrast to TTC where the collision is estimated assuming a single linear future trajectory for both vehicles A and B each, the risk of collision of the intended trajectory to be taken by vehicle B against all possible trajectories to be taken by vehicle A is evaluated. The risk value will then be a weighted combination of the single intended trajectory of vehicle B against the possible trajectories for vehicle A. The weights are assigned according to the probabilistic model for the future evolution of vehicle A behaviors.

1.3 Organization

The architecture and methods used for collision risk estimation can be found in [Sect. 2](#). Experimental results from Monte Carlo simulation and a realistic driving simulator environment can be found in [Sects. 3.1](#) and [3.2](#) respectively. This chapter ends with a conclusion in [Sect. 4](#).

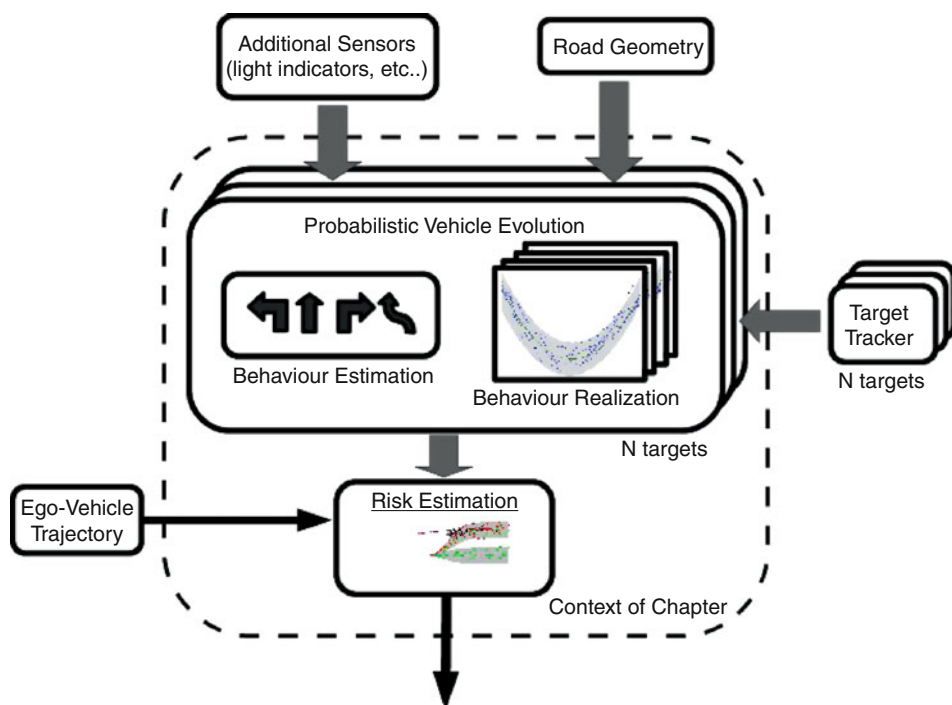
2 Collision Risk Estimation

The architecture of the system is presented in [Sect. 2.1](#). The probabilistic vehicle evolution model consists of two sub-modules: behavior estimation and realizations of behaviors. The probabilistic vehicle evolution model is then used in estimating the collision risk. They will be presented in [Sects. 2.2](#), [2.3](#) and [2.5](#) respectively.

2.1 Overall Architecture

[Figure 57.3](#) provides a global view of how the various components fit in within the global context. The problem is decomposed into sub-modules contained within the dotted box:

1. *Driving Behavior Recognition*: The aim of behavior recognition is to estimate the probability that a vehicle is executing one of the feasible behaviors. For example, it might give a probability value $P(\text{turn_left})$ that represents the probability that the vehicle observed will perform a turn left maneuver. As mentioned previously, behaviors are high level representations of road structure which contain semantics. The probability distribution over behaviors is performed by a Hidden Markov Model



■ Fig. 57.3

Overall view of how the risk estimation module fits in. Source: (TAY 2009)

(HMM). The current model has four behaviors; going straight, turning left, turning right, and overtaking. These will be described in greater detail in ▶ Sect. 2.2.

2. *Driving Behavior Realization*: When evaluating the risk of collision, it has to be performed geometrically. Driving behavior realization is represented as GPs which is a probabilistic representation of the possible future evolution of a vehicle given its behavior. The adaptation of GP according to the behavior is performed using geometrical transformation known as the Least Squares Conformal Map (LSCM). All relevant details will be described in ▶ Sect. 2.3.
3. *Evaluation of Risk*: A complete probabilistic model of the possible future evolution of a vehicle is given by the probability distribution over behaviors from *driving behavior recognition* and *driving behavior realization*. The risk of collision can be calculated based on this complete probabilistic model.


In general, the output of the risk of collision can be encapsulated under the intuitive notion of “risk of collision in the next few seconds.” However, its precise mathematical definition is highly dependent on the application. The model for risk estimation is compatible with a variety of risk estimation metrics according to the needs of applications. It will be described in detail in ▶ Sect. 2.5.

2.2 Behavior Recognition and Modeling

The aim of behavior recognition is to assign a label and a probability measure to sequential data. In this context, the sequential data received are the observations coming from the sensors. Examples of sensor values are distance to lane borders, signaling lights, or whether it is near an intersection, etc. However, the desired output is the probability values over behaviors. As such, the behaviors are hidden variables. There are a variety of related models for solving the problem assigning labels to sequences.

A well-known probabilistic model for inferring behaviors, based on sequential observations is the Hidden Markov Model (HMM) (Rabiner 1989). Extensions of the HMM includes Parametrized-HMM (Wilson and Bobick 1998), Entropic HMM (Brand and Kettnaker 2000), Variable-length HMM (Galata et al. 2001), Coupled HMM (Brand et al. 1997), and Structured HMM (Hongeng et al. 2000). These models extends the standard HMM for modelling complex activities and interactions.

This section presents a layered approach to model and estimate behaviors of vehicles under normal traffic conditions, as a means to the end within the context of estimating the risk of collision. The layered HMM (Oliver et al. 2002) decomposes the parameter space such that the robustness of the system is enhanced with the reduction of training and tuning requirements. Its architecture is very suitably applied to vehicle behavior modeling. Each layer contains a direct semantic equivalence which can be directly modeled.

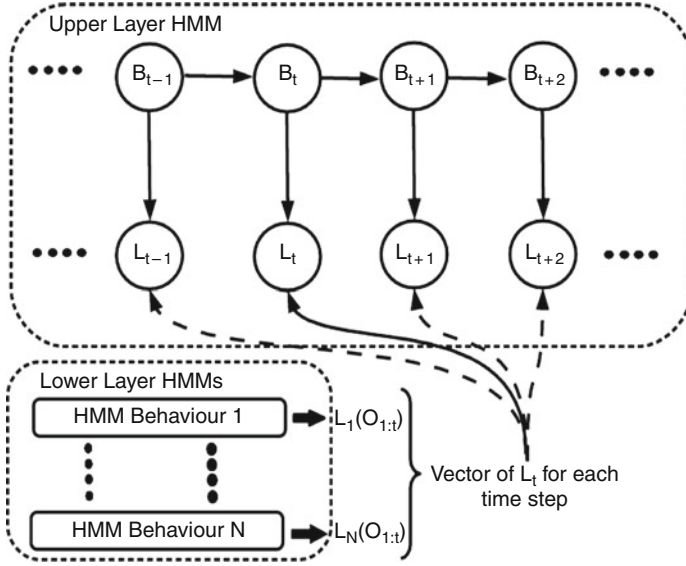
Behavior is modeled in two layers. Each layer consists of one or more HMMs. The upper layer is a single HMM where its hidden states represent behaviors at a high level, such as overtaking, turning left, turning right, or going straight. For each hidden State or behavior in the upper layer HMM, there is a corresponding HMM in the lower layer which represents the sequence of finer state transitions of a single behavior.  Figure 57.4 shows the schema for the layered HMM.

In this model, the following hidden state semantics in the lower layer HMMs for each of the following behaviors of the higher layer HMM are defined as follows:

- *Go Straight (one hidden state)*: go forward.
- *Overtake (four hidden states)*: lane change, accelerate (while overtaking vehicle), lane change back to original lane, resume normal speed.
- *Turn left/right (three hidden states)*: Decelerate before turn, execute turn, resume normal speed.

For purposes of inferring behavior of vehicles in the current context, it is desirable to maintain a probability distribution over the behaviors represented by the hidden states of the HMM in the upper layer. Observations made on vehicles coming from sensors interact with the HMM in the lower layer and information is then propagated up to the upper layer. In the lower layer, there is a corresponding HMM for each higher level behavior description. Each HMM in the lower layer, indexed by h , updates its current state by:

$$P(S_{t,h} O_{1:t}) \propto P(O_t | S_{t,h}) \sum_{S_{t-1,h}} P(S_{t-1,h}) P(S_{t,h} | S_{t-1,h}) \quad (57.1)$$



■ Fig. 57.4

Layered HMM. Each lower layer HMM's likelihood is computed and serves as the upper layer HMM's observation. Source: (TAY 2009)

where probabilistic variables O_t corresponds to observations at time t and $S_{t,h}$ is the variable for the hidden state of HMM h at time t . For each HMM h in the lower layer, its observation likelihood, $L_h(O_{1:t})$, can be computed:

$$L_h(O_{1:t}) = \sum_{S_{t,h}} P(S_{t,h} O_{1:t}) \quad (57.2)$$

Each of the observation likelihoods $L_h(O_{1:t})$ is the “observation” for the HMM of the upper layer. The inference of the upper level behaviors takes a similar form:

$$P(B_t | O_{1:t}) = P(O_{1:t} | B_t) \sum_{B_{t-1}} P(B_{t-1}) P(B_t | B_{t-1}) \quad (57.3)$$

$$= L_{B_t}(O_{1:t}) \sum_{B_{t-1}} P(B_{t-1}) P(B_t | B_{t-1}) \quad (57.4)$$

Where B_t is the hidden state variable of the upper level HMM at time t . $P(B_t | B_{t-1})$ is the upper level behavior transition matrix. Most of the time, it is reasonable to assume that a change in higher level behavior occurs more often after the end of the lower level behavior sequence, rather than in the middle of the lower level behavior sequence. An example is when a vehicle is executing the high level behavior of overtaking. A high level behavior of overtaking consists of lower level behaviors such as lane changing, accelerating past the other vehicle, return to original lane, and resuming normal speed. Chances of a vehicle changing high level behavior from overtaking to turning left, when the vehicle is at the lower level behavior of lane changing, are lower.

To take into account these effects, there are two different transition matrixes for the high level behavior. One transition matrix corresponds to the behavior transition when the lower level behaviors are completely performed (\mathbf{T}_{final}). Another transition matrix, $\mathbf{T}_{not-final}$ corresponds to the other case where lower level behaviors are not completely performed. Hence, the higher level behavior transition matrix can be calculated as a function of lower level states:

$$P(B_t|B_{t-1}) = \sum_{S_t, B_{t-1}} P(S_t, B_{t-1}) P(B_t|S_t, B_{t-1}) \quad (57.5)$$

where S_t, B_{t-1} is the state at time t of the HMM at the lower level, corresponding to the previous behavior B_{t-1} . $P(B_t|S_t, B_{t-1})$ is by definition:

$$P(B_t|S_t, B_{t-1}) = \begin{cases} \mathbf{T}_{final} & S_t, B_{t-1} \text{ is a final state} \\ \mathbf{T}_{not-final} & \text{otherwise} \end{cases} \quad (57.6)$$

At each time step, the probability distributions over high level behaviors $P(B_t|O_{1:t})$ are maintained iteratively. This will be used in the estimation of risk in [Sect. 2.5](#). The layered HMM is updated as follows in each time step:

2.3 Realizations of Behaviors

A behavior is an abstract representation of the motion of a vehicle. A probability distribution over the physical realization of the vehicle motion given its behavior is indispensable to the estimation of risk. The probability distribution over the physical realization of future vehicle motion is modeled using a GP.

Algorithm 1: Layered HMM Updates

Input: Current observation O_t
Output: $P(B_t|O_{1:t})$

```

1 foreach Lower layer HMM  $h$  do
2   | Update  $P(S_{t,h}|O_{1:t})$  (eqn. 1);
3   | Calculate log-likelihood  $L_h(O_{1:t})$  (eqn. 2);
4 end
5 Update upper layer HMM  $P(B_t|O_{1:t})$  (eqn. 4);
```

Recalling that the GP represents the normal driving routine where a driver approximately follows the lane and does not drift too far off to the left and right. On a straight road, this can be trivially represented with a GP where the mean of the GP corresponds to the middle of the lane (see [Fig. 57.5](#)).

A compact representation from the point of view of GPs does not involve learning separate GPs for the entire road network. To resolve cases where there are variations in curvature of lanes or for behaviors such as turning left or right, a procedure of adapting, what is referred to as a *canonical GP*, to the respective situations.

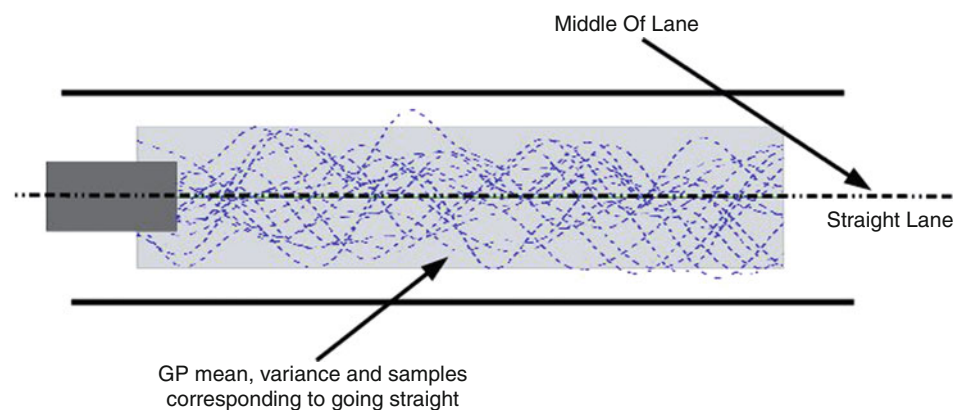
A *canonical GP* corresponds pictorially to [Fig. 57.5](#) where it is the GP corresponding to a vehicle traveling along a perfect straight stretch of road. The *canonical GP* serves as a basis from which it will be deformed to fit the situation required. The advantage of doing so is the compact and flexible representation of the possible lane geometry. Furthermore, a single GP can be calculated once and then reused for the different situations, thus gaining in speed and computation.

When nonlinear situations are encountered, a deformation will be performed on the *canonical GP* to fit the geometry of the lane. An example is shown in [Fig. 57.6](#) where the lane has a nonzero curvature.

2.3.1 Gaussian Process Deformation Model

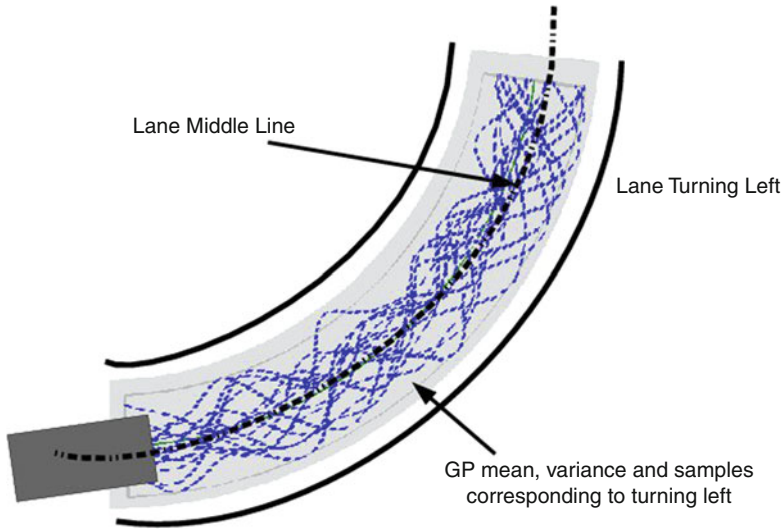
The aim of GP deformation is to adapt the *canonical GP* to the geometry of the lane. A natural way of looking at the adapted GP is to view the adapted GP as the same *canonical GP* defined in curvilinear coordinates. Hence, the problem of adapting the GP can be formulated as the invertible transformation, $\mathcal{U} : (x, y) \mapsto (u, v)$, mapping each single point of the canonical GP (x, y) defined in Cartesian coordinates to a single point (u, v) in curvilinear coordinates. \mathcal{U} is a one-to-one mapping and \mathcal{U}^{-1} exists (see [Fig. 57.7](#)).

Curvilinear coordinates appears in many engineering problems such as computational fluid dynamics or electromagnetics where a grid based on the curvilinear coordinates is used to solve partial differential equations numerically. The methods employed in these domains are not only computationally expensive but require the specification of the boundary. A common technique for the construction of curvilinear coordinates is *conformal mapping*.



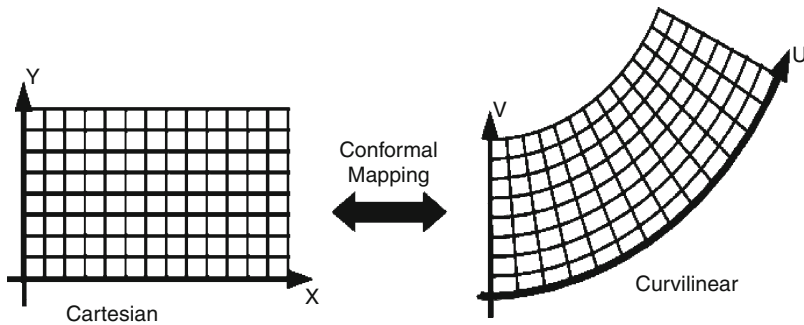
■ Fig. 57.5

Trivial example of the GP model for a perfectly straight lane. Source: (TAY 2009)



■ Fig. 57.6

Example of the deformed GP model for a lane turning left. Source: (TAY 2009)



■ Fig. 57.7

Invertible conformal map. Source: (TAY 2009)

Definition 2.1

A *conformal map* is a function of complex variables $\mathcal{U} : (x, y) \mapsto (u(x, y), v(x, y))$ which is analytic in the neighborhood of the open set containing (x, y) . Analytic functions are known to satisfy the Cauchy–Riemann equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \quad (57.7)$$

By differentiating [Eq. 57.7](#) with respect to x and y , and vice versa, the Laplace equation is obtained:

$$\Delta u = 0, \quad \Delta v = 0 \quad (57.8)$$

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplace operator. Since the mapping satisfies the Laplace equation, it is also known as a harmonic mapping.

2.3.2 Property

A conformal map produces smooth and invertible mappings of coordinate grids which minimizes distortion at a local level. It is a good candidate for the deformation of GP as not only it is smooth, it has an inverse mapping which is essential for performing prediction within the canonical GP frame. Furthermore, the local deformation is minimal as the Jacobian of \mathcal{U} is everywhere a rotation and scaling matrix.

2.3.3 Implementation Issues

Conformal mapping was originally defined in the continuous domain but is computationally demanding. Discrete conformal mapping techniques that approximate this process perform piecewise linear mappings between triangles of the mesh. Most current methods approximate the conformal map by discretization of the Laplace operator at the vertexes of the mesh triangles. Such solutions usually require the specification of boundary conditions (Pinkall et al. 1993; Eck et al. 1995).

2.3.4 Chosen Implementation

In the context of the problem, the specification of the boundary is unnecessary. The lane boundary is implicitly defined by the curve representing the middle of the lane, and the width of the lane. Furthermore, specifying the boundary of the lanes is not straightforward especially in portions of lanes with high curvature. Ideally, it is simply sufficient to be able to generate the curvilinear coordinates based purely on the line or curve representing the middle of the lane and the width of the lane.

A dual approach which avoids the specification of all boundary solution, the LSCM, was proposed (Lvy et al. 2002). Instead of discretizing the Laplace operator at the vertexes of the triangulation, LSCM proposes to adhere as much as possible the conformality condition in each of the triangles of the triangulation, reducing the problem into an unconstrained quadratic minimization problem which can be efficiently solved numerically.

Rewriting the conformal map \mathcal{U} in \mathbb{C} where $\mathcal{U} = u + iv$, the Cauchy–Riemann conditions can be written equivalent as:

$$\frac{\partial \mathcal{U}}{\partial x} + i \frac{\partial \mathcal{U}}{\partial y} = 0 \quad (57.9)$$

The LSCM seeks to minimize the violation of the conformality criterion of [Eq. 57.9](#) on all triangles of the triangulation:

$$C(\mathcal{T}) = \sum_{T \in \mathcal{T}} \int_T \left| \frac{\partial \mathcal{U}}{\partial x} + i \frac{\partial \mathcal{U}}{\partial y} \right|^2 dA \quad (57.10)$$

$$= \sum_{T \in \mathcal{T}} \left| \frac{\partial \mathcal{U}}{\partial x} + i \frac{\partial \mathcal{U}}{\partial y} \right|^2 A_T \quad (57.11)$$

where \mathcal{T} is the set of triangles in the triangulation, and A_T is the area of the triangle. Consider the mapping of a single triangle in complex space with points $\{(x_i, y_i)\}_{i=1..3}$ via \mathcal{U} giving $\{(u(x_i, y_i), v(x_i, y_i))\}_{i=1..3}$ respectively. For a conformal mapping, the Jacobian is everywhere a scalar times rotation matrix, the mapping for a single triangle can be represented as a rotation and translation:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} \partial u / \partial x & \partial u / \partial y \\ \partial v / \partial x & \partial v / \partial y \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \quad (57.12)$$

The rotation matrix can be obtained by solving a system of linear equations given six unknowns and the three point correspondences. Hence the gradient vector of $u(x, y)$ is:

$$\begin{pmatrix} \partial u / \partial x \\ \partial u / \partial y \end{pmatrix} = \frac{1}{D} \begin{pmatrix} y_2 - y_3 & y_3 - y_1 & y_1 - y_2 \\ x_3 - x_2 & x_1 - x_3 & x_2 - x_1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad (57.13)$$

where D is twice the area of the triangle ($D = (x_1 y_2 - y_1 x_2) + (x_2 y_3 - y_2 x_3) + (x_3 y_1 - y_3 x_1)$). The gradient vector of $v(x, y)$ is similar. Hence the Cauchy–Riemann [Eq. 57.9](#) can be written as:

$$\frac{\partial \mathcal{U}}{\partial x} + i \frac{\partial \mathcal{U}}{\partial y} = \frac{i}{D} (W_1 \ W_2 \ W_3) (U_1 \ U_2 \ U_3)^T \quad (57.14)$$

where

$$W_1 = (x_3 - x_2) + i(y_3 - y_2) \quad (57.15)$$

$$W_2 = (x_1 - x_3) + i(y_1 - y_3) \quad (57.16)$$

$$W_3 = (x_2 - x_1) + i(y_2 - y_1) \quad (57.17)$$

$$U_i = u(x_i, y_i) + iv(x_i, y_i) \quad (57.18)$$

The minimization of the violation of the conformality criterion in [Eq. 57.11](#) can be written in its discrete form:

$$C(\mathcal{T}) = \sum_{T \in \mathcal{T}} C(T) \quad (57.19)$$

$$= \sum_{T \in \mathcal{T}} \frac{1}{D_T} \left| (W_{1,T} \ W_{2,T} \ W_{3,T}) (U_{1,T} \ U_{2,T} \ U_{3,T})^T \right|^2 \quad (57.20)$$

$$= (\mathcal{M}\mathbf{U})^* (\mathcal{M}\mathbf{U}) \quad (57.21)$$

where $W_{i,T}$ and $U_{i,T}$ are the values of W_i and U_i corresponding to triangle T respectively. $\mathbf{U} = (U_1, \dots, U_n)$ for n vertexes of the triangulated mesh. \mathcal{M} is a sparse matrix, of dimension $n' \times n$ where rows correspond to triangles and columns its vertexes. Each entry of \mathcal{M} , m_{ij} , contains the values:

$$m_{ij} = \begin{cases} \frac{W_{j,T_i}}{\sqrt{d_{T_i}}} & \text{if vertex } j \text{ belongs to triangle } T_i \\ 0 & \text{otherwise} \end{cases} \quad (57.22)$$

Usually, a set of points, $p_i = (x_i, y_i)$, which makes up the vertexes of the triangular mesh in the original space are given. A subset of points U_i are fixed *a priori* which represents the user determined positions $\mathcal{U}(p_i) = U_i$. LSCM then computes the coordinates of the remaining *free points* given the user specified *fixed points*. Denoting the vector of free points as \mathbf{U}_f and vector of fixed points as \mathbf{U}_p , the vector \mathbf{U} and matrix \mathcal{M} can be similarly decomposed in the following way:

$$\mathbf{U} = (\mathbf{U}_f^T \ \mathbf{U}_p^T) \quad (57.23)$$

$$\mathcal{M} = (\mathcal{M}_f \ \mathcal{M}_p) \quad (57.24)$$

where \mathcal{M}_f and \mathcal{M}_p are block matrices of dimensions $n' \times (n - p)$ and $n' \times p$ respectively. The equation to be minimized ([Eq. 57.21](#)) can be now written as:

$$C(\mathcal{T}) = \|\mathcal{M}_f \mathbf{U}_f + \mathcal{M}_p \mathbf{U}_p\|^2 \quad (57.25)$$

[Equation 57.25](#) can be solved using the Moore–Penrose pseudoinverse:

$$\mathbf{U}_f = (\mathcal{M}_f^* \mathcal{M}_f)^{-1} \mathcal{M}_f^* (\mathcal{M}_p \mathbf{U}_p) \quad (57.26)$$

However, for large number of *free points*, the matrix $(\mathcal{M}_f^* \mathcal{M}_f)$ which is of size $(n - p) \times (n - p)$ will be large and involves a large number of multiplications. Furthermore, inversion of matrices has complexity $O(n^3)$. A faster method will be to use the conjugate gradient (Hestenes and Stiefel 1952) to perform the inversion which reduces it to $O(n^2)$.

2.4 Predicting Vehicle Motion

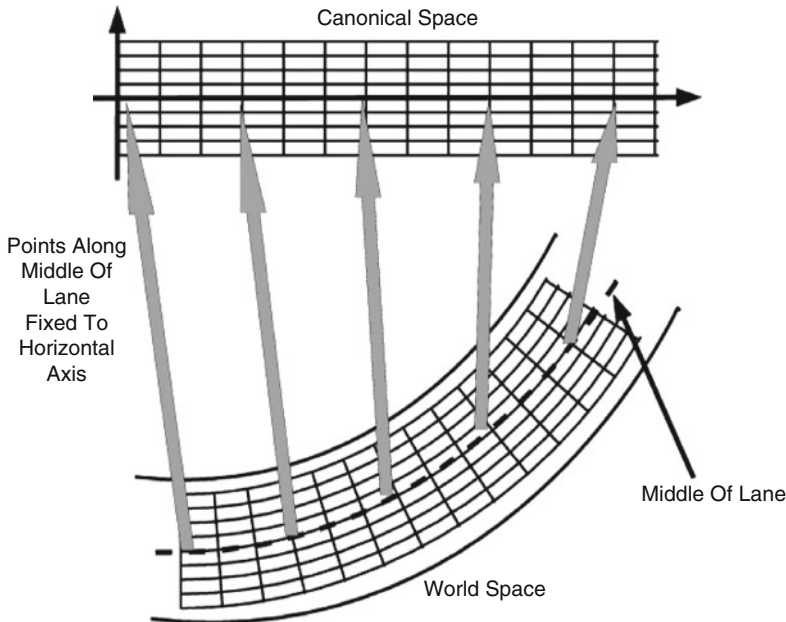
The previous [Sect. 2.3.1](#) described an isomorphic mapping between the GP adapted to the road geometry and the *canonical GP*. This section presents the procedure on using the mapping for predicting vehicle motion.

An informed prediction on vehicle motion requires the observation of the current and past states. At every time instance t , a temporally ordered sequence of current and past observations $O = \{O_t, O_{t-1}, \dots, O_{t-K}\}$ is maintained where each observation $O_{t-k} = (x_{t-k}, y_{t-k})$ is a vector containing the positions of the vehicle. The observations are then transformed via LSCM to *canonical GP* coordinates, where the future motion can be inferred. The probability distribution over future motion is then transformed back to real-world coordinates. Each stage of the procedure for predicting vehicle motion is detailed.

2.4.1 Conformal Transformation Between World Space and Canonical Space

The mapping between world space and canonical space (the space where the *canonical GP* resides within) is discretized and represented as an isomorphic mapping between two meshes. This is done via LSCM (see Sect. 2.3.1).

Obtaining the mapping requires the specification of a certain number of *fixed points* and its mapped coordinates. The *fixed points* are deterministically chosen; a discretized set of points lying along the middle of the lane, each corresponding to a point along the horizontal axis of the *canonical GP* frame (see Fig. 57.8).



■ Fig. 57.8

Conformal transformation between canonical space and world space. Fat arrows show the fixed points where points along middle of lane in world space are mapped to horizontal axis of canonical space. Source: (TAY 2009)

The middle of the lane is described as a poly-line. Such information can come from processed sensor data such as the lidar or camera. To determine the set of *fixed points*, the initial point is chosen by perpendicularly projecting the oldest observation, $O_l = \min_t O_t$, $O_t \in O$ on to the poly-line to obtain P_0 . From P_0 , a sequence of points, $P = \{P_1, \dots, P_N\}$, along the poly-line are obtained such that for any two consecutive points, $\|P_i - P_{i+1}\|_2 = d$, where d is a constant. Each point P_i is associated with a vertex of the mesh in world space and is mapped to a vertex of the mesh in canonical space where $\mathcal{U}(Q_i) = P_i$ where Q_i has the coordinates $(d^* i, 0.0)$ in canonical space.

The transformation is fully defined by having each vertex of the mesh in canonical space mapped to a vertex of the mesh in world space and vice versa. The vertexes of the mesh in canonical space are arranged in a grid and its points are known *a priori*. Let P' be the set of vertexes in world space with free coordinates (unspecified so far). P' can then be obtained by solving \blacklozenge Eq. 57.26 where $\mathcal{U}_f = P'$ and $\mathcal{U}_p = P$.

2.4.2 Inferring Probability Distribution on Future Motion

The observations in world coordinates have to be mapped to canonical space before inference on future motion can be performed. The *LSCM* gives the discrete piecewise affine mapping between the two spaces. Observations in world coordinates can be mapped to canonical space via $\mathcal{U}^{-1}(O_i) = (x_i, y_i)$.

The mapping \mathcal{U} is discretized and manifests in the form of a mesh. $\mathcal{U}^{-1}(O_i)$ can be calculated by first locating the mesh triangle which contains O_i in the world space mesh, and then transform O_i back to the corresponding mesh triangle in canonical space by calculating the corresponding barycentric coordinates.

The mapping of the past n observations of vehicle positions in world coordinates gives a set of values $\{(x_i, y_i)\}_{i=1}^n$ in canonical space. The probability distribution over future motion of the observed vehicle thus corresponds to the probability distribution given by the GP (\blacklozenge Fig. 57.9):

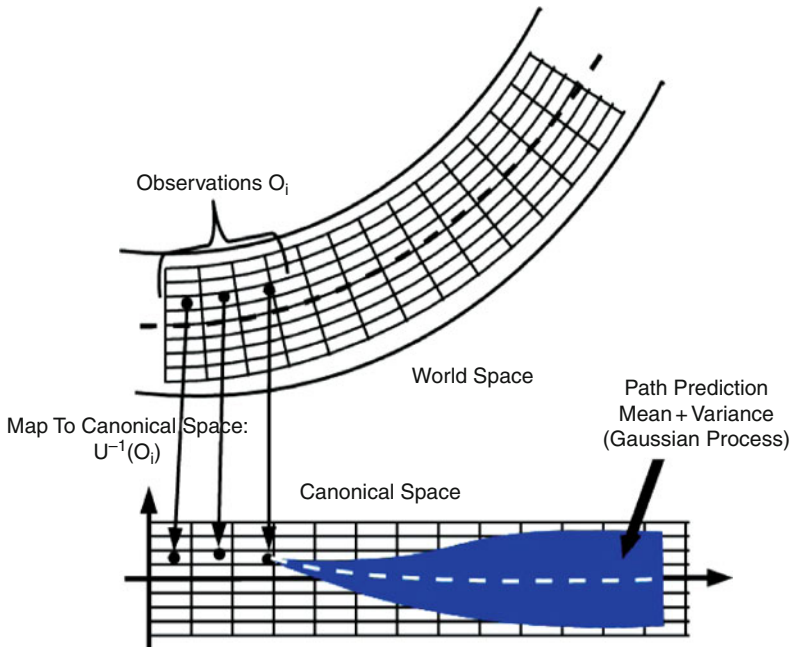
$$P(Y_* | X_*, X, Y) = \mathcal{GP}(\mu_{Y_*}, \Sigma_{Y_*}) \quad (57.27)$$

$$\mu_{Y_*} = K(X_*, X) [K(X, X) + \sigma^2 \mathbf{I}]^{-1} Y \quad (57.28)$$

$$\Sigma_{Y_*} = K(X_*, X_*) - K(X_*, X) [K(X, X) + \sigma^2 \mathbf{I}]^{-1} K(X, X_*) \quad (57.29)$$

where $X = (x_1, \dots, x_n)^T$, $Y = (y_1, \dots, y_n)^T$ are the observations. $X_* = (x_1^*, \dots, x_K^*)$ is the vector of x values for which the predicted values which is represented by $Y_* = (y_1^*, \dots, y_K^*)$ and each $x_i^* > \max X$. The covariance function used is the squared exponential:

$$k(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right) \quad (57.30)$$



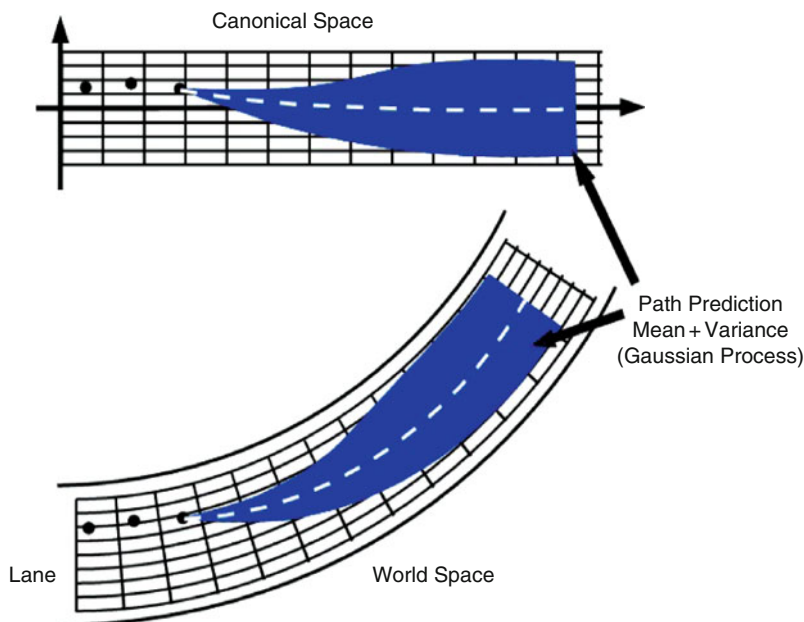
■ Fig. 57.9

Observations are mapped into canonical space before conditioning Gaussian Process on observations to obtain probability distribution over future motion. Source: (TAY 2009)

2.4.3 Mapping Back to Real-World Coordinates

The probability distribution over future motion is a Gaussian Process in canonical space specified by [Eq. 57.29](#) and has to be mapped back to world space in order to evaluate the risk of collision. However, the conformal transformation of a Gaussian Process is not trivially defined. Fortunately, sampling from a Gaussian distribution is trivial. A Monte Carlo approximation of the distribution by first sampling from $P(Y_* | X_*, X, Y)$ is used. The samples will later be used to evaluate the risk ([Sect. 2.5](#)). Intuitively each sample is a possible realization of the future vehicle motion, represented as a sequence of position values $S_i = ((x_{i,1}^*, y_{i,1}^*), \dots, (x_{i,K}^*, y_{i,K}^*))$. As the samples are in canonical space, it is transformed back to world space via LSCM.

The procedure for the transformation is similar to that of mapping observations in world space to canonical space, except that the mapping is in the inverse direction. Each point of sample S_i is mapped by locating the mesh triangle in canonical space containing the point and mapped to the corresponding mesh triangle in world space by calculating the barycentric coordinates ([Fig. 57.10](#)).



■ Fig. 57.10

The *canonical GP* is transformed back to world space. Shaded regions display the mean and variance of the GP. Transformation can be approximated by sampling from the GP in canonical space before transforming the samples. Source: (TAY 2009)

2.5 Evaluation of Risk

The layered HMM approach (● Sect. 2.2) assigns a probability distribution over behaviors at each time instance. And for each behavior, a Gaussian Process gives the probability distribution over its physical realization. Because the behavioral semantics are propagated from the layered HMM right down to the physical level, it is now possible to assign semantics to risk values as well.

It is important to note that the definition of risk can take a variety of forms, which is largely dependent on how the risk output is going to be used. A risk scalar value might be sufficient for a crash warning system, or an application might require the risk values against each individual vehicle in the scene. The application scope using such risk values can be classified into two different categories.

The first category of applications involves a varying degree of vehicle control where risk values can be used to drive an autonomous vehicle, or simply to take control of a vehicle to avert the vehicle away from danger momentarily.

The second category of applications are passive in nature where no feedback into the control loop is involved. An example is a passive driving assistance system which warns drivers of possible danger ahead. The risk evaluation is illustrated in a generic bottom-up manner with varying risk semantic in the following:

2.5.1 Risk of Trajectory Considering Behavior of One Vehicle Only

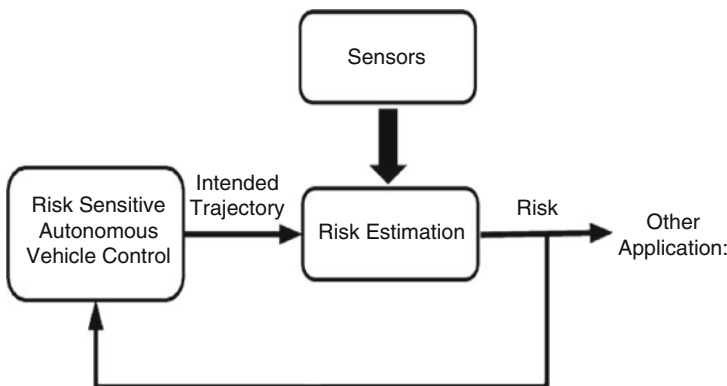
Suppose a simple example of an autonomous vehicle navigating through a dynamic environment, avoiding collisions with the moving entities in the environment. Such autonomous vehicles usually have a feedback control or navigation system. Apart from low level control issues such as trajectory following, a control module which takes into account the risk of collision is considered as well. It is not difficult to imagine that this control module works by evaluating a set of potential trajectories to be taken by the autonomous vehicle and that the autonomous vehicle will choose the trajectory with the lowest risk within the considered set (► Fig. 57.11).

In this case, the risk of a single considered trajectory is calculated. In a scene, there might be several vehicles present. Consider the simple case of only one vehicle present, vehicle V_1 , (excluding autonomous vehicle, V_A). The risk of a trajectory considered by V_A , trajectory T_A , against behavior b of vehicle V_1 is given by:

$$P(C|T_A B_{V_1} V_1) = \sum_{T_{V_1}} P(C|T_A T_{V_1} B_{V_1} V_1) P(T_{V_1}|B_{V_1} V_1) \quad (57.31)$$

Where C is a probabilistic boolean variable indicating if there is a collision, B_{V_1} is the variable corresponding to the behaviors for vehicle V_1 , described by the hidden states of the upper layer HMM. T_A and T_{V_1} are the trajectories of V_A and V_1 respectively. $P(T_{V_1}|B_{V_1} V_1)$ is the physical realization of behavior B_{V_1} and thus is represented by the trajectories sampled from the Gaussian Process mentioned previously in ► Sect. 2.4. $P(C|T_A T_{V_1} B_{V_1} V_1)$ evaluates whether there is a collision between trajectories T_A and T_{V_1} .

In reality T_A and T_{V_1} is a list of points describing the path, i.e., $T_i = (P_1^i, \dots, P_k^i)$, $P_j^i = (x_j^i, y_j^i)$. However, the speed and accelerations of V_1 are available from a target tracker. Speed and acceleration on V_A can be obtained from its proprioceptive sensors. A constant acceleration model is used to compute if there is a collision. Based on the velocity and acceleration of V_A and V_1 , its positions along trajectories T_A and T_i



■ Fig. 57.11

Architecture of a simple risk sensitive control of an autonomous vehicle. Source: (TAY 2009)

respectively can be easily calculated by linearly interpolation along the list of positions describing T_A and T_i . These positions are calculated in discrete time steps and at each time step, a collision detection is performed.

A subtle point when performing collision detection is that the geometry of the vehicles has a significant influence in the final risk values. It might be easy to think that a collision detection based on the L2-distance between two coordinates, coupled with the averaging effects over the sampled trajectories, will yield a proper estimate. It has been observed in experiments that this is not so. Vehicles passing by each other on adjacent lanes consistently give a false high collision probability. On the other hand reducing the L2-distance threshold will give overly optimistic estimates of collision probability when one vehicle is behind the other. This is a consequence of the geometry of the vehicle (rectangular) where the length of the vehicle is longer than its width. The assumption that all vehicles are of the same length and width was made. The collision detection between two rectangles representing the geometry of the vehicles is performed by searching for an axis separating the two rectangles. The following pseudo-code (algorithm 2) gives a summary:

Algorithm 2: Evaluation of $P(C T_A B_{V_i} V_i)$	
	Input: Trajectory T_A for vehicle V_A
	Output: $P(C T_A B_{V_i} V_i)$
1	ColCount = 0.0;
2	foreach <i>Sampled path</i> $T_{V_i} \sim P(T_{V_i} B_{V_i} V_i)$ do
3	foreach <i>Discretized time step</i> $t = \text{step} * \Delta t$ do
4	X_A = Position of V_A at time t along polyline T_A ;
5	X_1 = Position of V_i at time t along polyline T_{V_i} ;
6	Θ_A = Orientation of line segment of T_A containing X_A ;
7	Θ_1 = Orientation of line segment of T_{V_i} containing X_1 ;
8	R_A = Rectangle centered at X_A and angle Θ_A ;
9	R_1 = Rectangle centered at X_1 and angle Θ_1 ;
10	if <i>Separating axis exist between</i> R_A <i>and</i> R_1 then
11	ColCount = ColCount + 1.0;
12	end
13	end
14	end
15	return ColCount / Number of Samples Paths;

2.5.2 Risk of trajectory against one vehicle with behaviors aggregated

The risk of a trajectory against another vehicle, can be obtained by aggregating the risk previously (against one behavior of another vehicle). The aggregation is essentially a weighted sum of $P(C|T_A B_{V_i} V_i)$ for each behavior B_{V_i} of vehicle V_i .

$$P(C|T_A V_i) = \sum_{B_{V_i}} P(C|T_A B_{V_i} V_i)P(B_{V_i}|V_i) \quad (57.32)$$

The weighted sum was performed against the term $P(B_{V_i})$ in [Eq. 57.32](#) and its values come from the layered HMM (see [Sect. 2.2](#), [Eq. 57.6](#)).

2.5.3 Aggregating Risk with Respect to All Vehicles

The risk of trajectory T_A when taking a single vehicle V_i into account is represented by $\mathcal{R} = P(C|T_A V_i)$. There are several possible choices for aggregating risk, which is largely dependent on how the aggregated risk value is going to be used or interpreted. The function of risk aggregation is a function of risk values with respect to all vehicles, i.e., $\mathcal{F}(\mathcal{R}_1, \dots, \mathcal{R}_N)$ for N vehicles in the scene:

- *Marginalizing over vehicles*: A direct way of aggregating risks will be to marginalize over the prior probabilities of the vehicles:

$$\begin{aligned}\mathcal{F}(\mathcal{R}_1, \dots, \mathcal{R}_N) &= P(C|T_A) \\ &= \sum_{V_i} P(C|T_A V_i)P(V_i) \\ &= \sum_{V_i} \mathcal{R}_i P(V_i)\end{aligned}\tag{57.33}$$

- The prior probabilities over vehicles, $P(V_i)$, in [Eq. 57.33](#) can come from an object recognition module which expresses the confidence that object V_i is a vehicle. Without any information, a uniform prior can be used instead and is equivalent to taking the average risk of all vehicles.
- *Maximum risk*: Marginalizing over vehicles might be under conservative in some cases. This is especially so when a single vehicle poses an imminent danger in a scene with many vehicles and the average gives a low estimate. In this case, taking the maximum risk value might represent the risk more accurately:

$$\mathcal{F}(\mathcal{R}_1, \dots, \mathcal{R}_N) = \max_{V_i} P(C|T_A V_i)\tag{57.34}$$

- *Temporally nearest risk*: The evaluation of collision risk, $P(C|T_A B_{V_1} V_1)$ (algorithm 2), does not explicitly take into account time. For example, the check for collision between T_A and sampled trajectory T_{V_i} in algorithm 2 only indicates if there is a collision in a certain time horizon in the future regardless of the length of horizon:

$$\text{Collide}(T_A, T_{V_i}) = \begin{cases} 1.0 & \text{If any collision exists in time horizon} \\ 0.0 & \text{otherwise} \end{cases}\tag{57.35}$$

- Incorporating time into risk evaluation is useful in certain cases. For applications such as crash warning, it is less probable that if the driver maintains the current acceleration, a crash 30 s in the future is unavoidable with probability 1.0. The drivers of the vehicles involved have reasonable time to react to the situation. In this case, it might be

desirable to express risk further ahead in time as having less “importance.” This can be taken into account by modifying algorithm 2 where the risk is weighted by a decreasing function with time:

$$Collide^*(T_A, T_{V_i}) = \begin{cases} \exp^{-\alpha t^2} & \text{If collision between } T_A \text{ and } T_{V_i} \\ 0.0 & \text{otherwise} \end{cases} \quad (57.36)$$

where t represents the amount of time before collision occurs and α is a constant which expresses the rate of risk decrease with time.

2.5.4 Risk Associated with Driving Behavior

So far, the risk value of a single trajectory T_A for an autonomous vehicle is calculated. For applications where risk values are passively used, especially when the driver is a human and not a computer program, it is less practical to evaluate the risk of only a single trajectory T_A . The alternative will be to evaluate risks associated with behavior or general risk value for the ego-vehicle (the vehicle for which the risk shall be evaluated for, but shall be named V_A still).

- *Behavior-related risk:* Instead of evaluating for a single T_A , the risk is evaluated for the collection of T_A associated with a behavior of the ego-vehicle. For example, to obtain the risk of a certain behavior of ego-vehicle, against another vehicle, V_i :

$$P(C|B_{V_A} V_i) = \sum_{T_A, B_{V_i}} P(C|T_A B_{V_i} B_{V_A} V_i) P(T_A|B_{V_A}) P(B_{V_i}|V_i) \quad (57.37)$$

where $P(T_A|B_{V_A})$ is the probability distribution over the future trajectory of the ego-vehicle with behavior B_{V_A} . $P(C|T_A B_{V_i} B_{V_A} V_i)$ is the collision risk of trajectory T_A against vehicle V_i with behavior B_{V_i} . The evaluation of this term is exactly the same as algorithm 2. Essentially, the algorithm for [Eq. 57.37](#) will be to sample an ego-vehicle trajectory $P(T_A|B_{V_A})$ and each sample is evaluated against the sampled trajectories of vehicle V_i across all behaviors:

Algorithm 3: Evaluation of $P(C|B_{V_A} V_i)$

Input: Ego-vehicle behaviour B_{V_A}
Output: $P(C|B_{V_A} V_i)$

```

1 ColCount = 0.0;
2 foreach Sampled Trajectory  $T_A \sim P(T_A|B_{V_A})$  do
3   | trajCol = Evaluate algorithm 2 with  $T_A$  as parameter;
4   | ColCount = ColCount + trajCol;
5 end
6 return ColCount / Number of sampled  $T_A$ ;
```

- **General risk value:** A risk value between vehicle V_A and V_i can be obtained from 37 by marginalization over the estimated behavior of the ego-vehicle:

$$P(C|V_i) = \sum_{B_{V_A}} P(C|B_{V_A} V_i) P(B_{V_A}) \quad (57.38)$$

$P(B_{V_A})$ is the distribution over the behavior of the ego-vehicle. This can be obtained by application of the layered HMM on the ego-vehicle.

Several examples of risk with different semantic are presented. The number of different ways of evaluating risk is combinatorial. Risk can be evaluated between trajectory samples, behaviors, vehicles, or all vehicles. This is to highlight the flexibility of the current system of using a HMM-based object in identifying behaviors coupled with the use of GP for behavior realizations, while taking the road geometry and topology into account.

3 Experiments

Two different experimental validations within the context of driving assistance were conducted:

- The first is based on Monte Carlo simulations to validate its accuracy and reliability under a variety of situations and scenarios.
- Due to the nature of the application, it is impossible to produce real life crash testing. However, experiments based on an elaborate human-driven scenario in a virtual environment were conducted and the results are presented. The experiments were conducted in collaboration with Toyota Motors Europe and ProBayes.

3.1 Monte Carlo Simulation Validation

A pertinent question when performing experimental validation of the estimation of the risk of collision is its reliability. As a large number of different scenarios are required, it is infeasible in practice. A better method is to experimentally evaluate the estimation of risk of collision by randomly generating many scenarios under a variety of situations. To this end, a Monte Carlo-based approach is adopted to sample the different scenarios for statistical evaluation.

The Monte Carlo simulations are performed over the space of different possible situations that can occur in a normal road traffic environment. The situations in this context include the different topology of the roads, the various configurations the vehicles in the scene are positioned, and its associated dynamics.

Each sample thus represents a scene with a number of vehicles. Within each sample, one vehicle is identified as the ego-vehicle where the estimation of the risk of collision will be performed for the ego-vehicle.

3.1.1 Experimental Setup

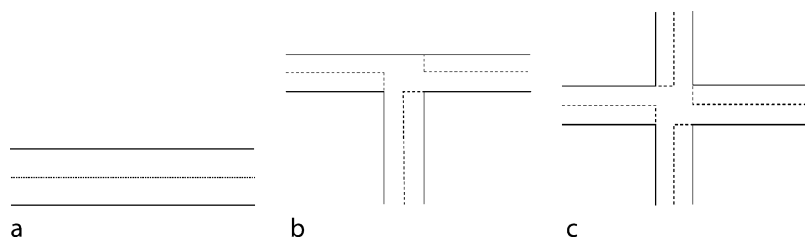
One advantage of Monte Carlo-based simulation is the ability to categorically analyze the algorithm under different specific situations. The situations are basically hierarchically organized according to the topology of the road, and the behavior of vehicle V_A which is the vehicle in question for which the risk of collision is to be estimated.

In the experimental setup, basically three basic road topology were identified, namely, the parallel road, T-junction, and cross-junction (see [Fig. 57.12](#)):

For each road topology, the situations can be artificially generated by generating a random number of vehicles in the scene. Each randomly sampled vehicle in the scene includes randomly chosen parameters such as the intended route (given indirectly from a randomly chosen behavior), its starting position, its velocity, and acceleration. One of the vehicles is designated as vehicle V_A for which the estimation of the risk of collision will be calculated. Each randomly sampled situation will then be simulated by evolving the random vehicles in the scene in time. At every time step, the risk of collision of the ego-vehicle is calculated and recorded.

In practice, the generation of the samples is performed in a hierarchical fashion (see [Fig. 57.13](#)). The upper level of the hierarchy represents the different road topology. The middle level represents the different behaviors for the ego-vehicle within each road topology. For example, the middle level will represent the situation where the ego-vehicle is going straight, turning left, and turning right for the road topology T-junction. The bottom level will then be each individual situation/samples. Additionally, a Gaussian noise is added along the samples to test its robustness.

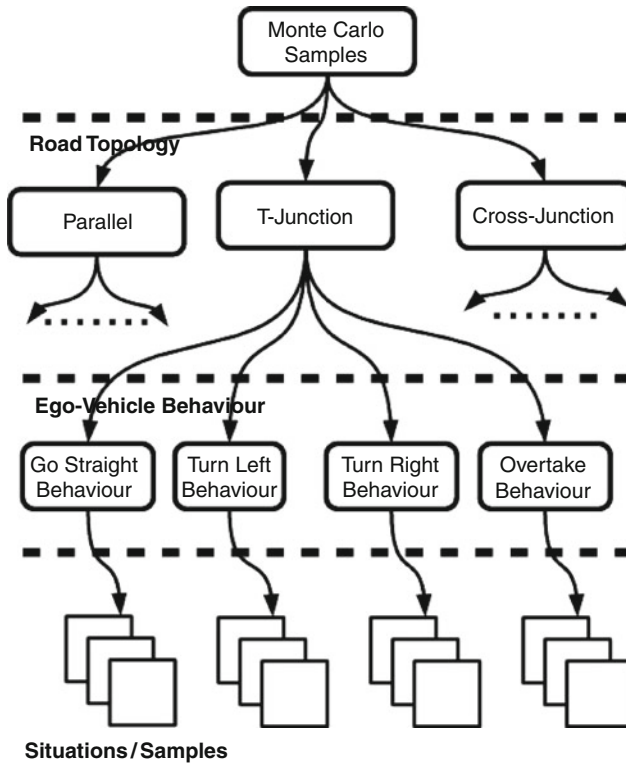
For the Monte Carlo experiments, the estimation of the behaviors, which in reality is to be estimated by the HMM, is assumed to be fully known. The reason for this is to evaluate the estimation of the risk of collision independently from the output coming from the layered HMM.



■ Fig. 57.12

Three different road topologies corresponding to Parallel, T-Junction and Cross-Junction.

Source: (TAY 2009)



■ Fig. 57.13

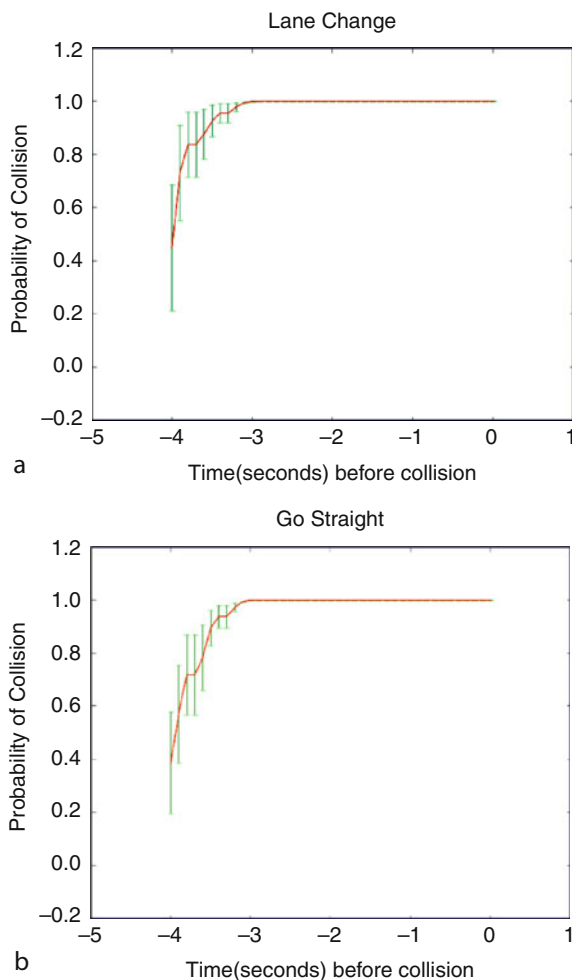
Organization of Monte Carlo simulation. Source: (TAY 2009)

3.1.2 Results

The results in this sub-section present the graph of the evolution of risk values 4 s before collision. Two hundred situations were randomly sampled from every combination of road topology and behavior for vehicle V_A . For each instance in time where there is a collision for the ego-vehicle, the values of the estimation of risk of collision for the 4 s duration before collision are recorded. The plot of the mean and variance of risk values 4 s before each collision can be obtained.

In the experiments, the risk evaluated for a vehicle V_A is the risk of its intended trajectory T_A . This risk value can be used as a feedback output to a risk sensitive vehicle control module, or it can be easily generalized to be a collision warning system, by evaluating different trajectories T_A as described in ▶ Sect. 2. The risk of trajectory T_A against another vehicle V_i is given by:

$$P(C|T_A V_i) = \sum_{B_{V_i}} P(C|T_A B_{V_i} V_i) P(B_{V_i}|V_i) \quad (57.39)$$

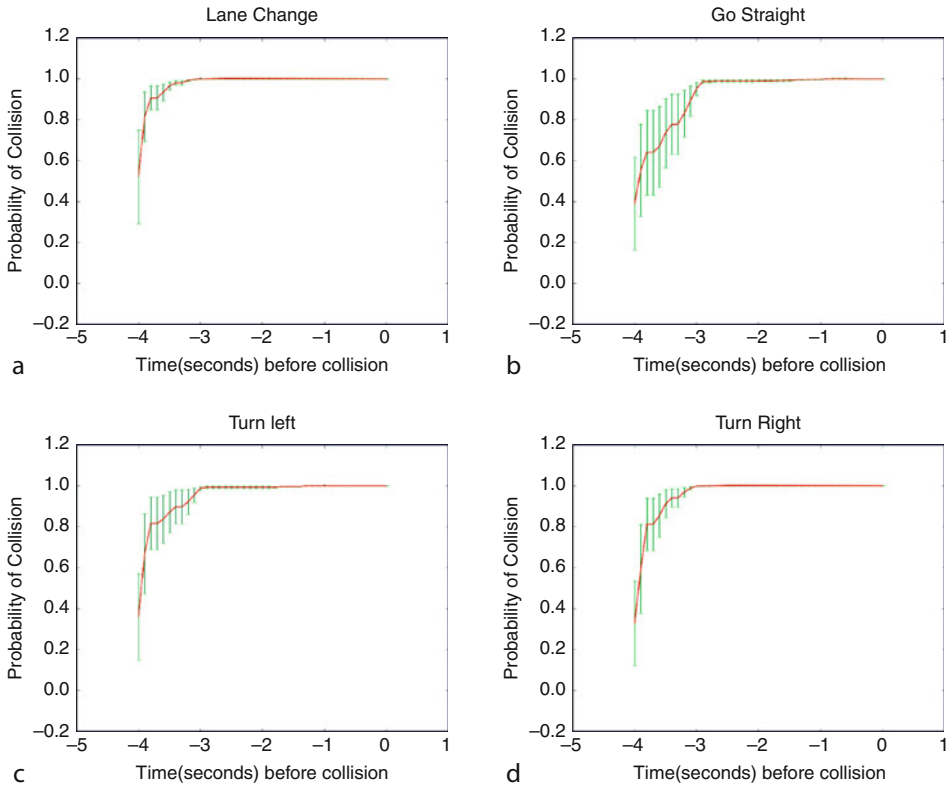


■ Fig. 57.14

Plot of mean and variance of risk estimation for parallel road. Source: (TAY 2009)

Where $P(B_{V_i}|V_i)$ originally refers to the probability distribution over behaviors from the layered HMM corresponding to vehicle V_i . However, to evaluate the risk estimation independently from the layered HMM, $P(B_{V_i}|V_i)$ in the Monte Carlo experiments is a Dirac distribution centered at the known behavior B_{V_i} for vehicle V_i .

The time horizon of risk evaluation is for 3 s. Technically, this means that the prediction of future motion trajectory is limited to a distance of the maximum speed multiplied by 3 s. Finally, the risk among all vehicles is aggregated by taking the maximum risk of all vehicles.



■ Fig. 57.15

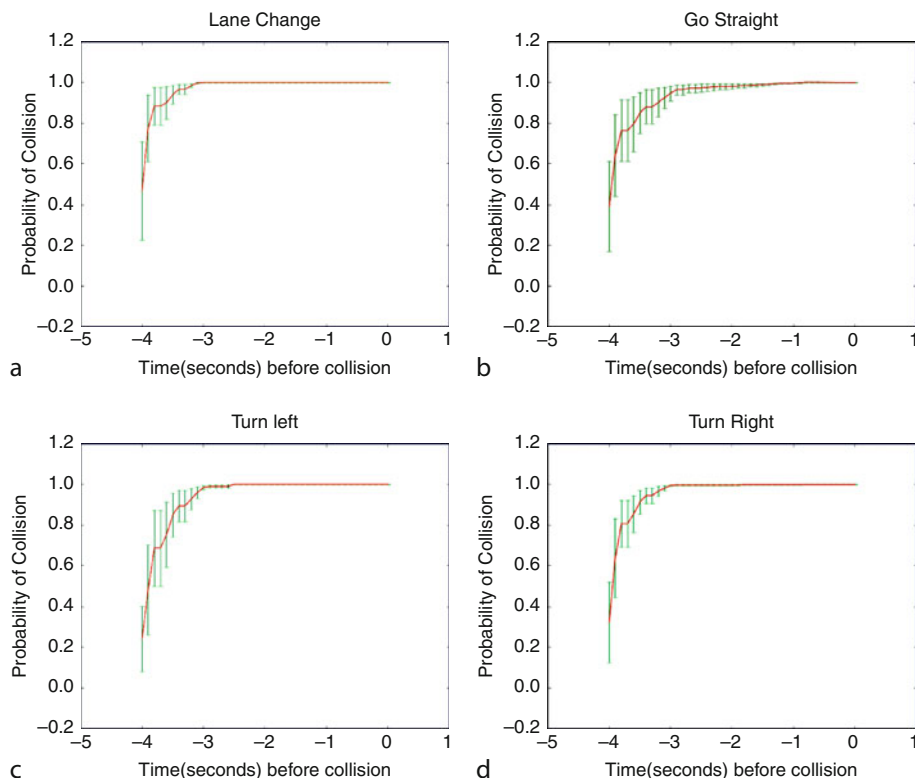
Plot of mean and variance of risk estimation for T-Junctions. Source: (TAY 2009)

► Figure 57.14 shows the mean and variance of the risk values 4 s before collision, along a parallel stretch of road. Each sub-figure displays the corresponding plots when vehicle V_A is executing a certain behavior.

► Figures 57.15 and ► 57.16 show the corresponding plots of risk means and variance in a T-junction and cross-junction respectively, along with its different behaviors for vehicle V_A in each of the T-junction and cross-junction.

It can be observed from all the graphs in ► Figs. 57.14–57.16 that the risk estimated 3 s before collision consistently achieves a value of approximately 1.0 across all graphs. As the risk value is a probability value, this means that the risk estimation for collision is consistent and reliable for the time horizon considered. However, these set of results are obtained knowing fully the behavior of the vehicles in the scene, thus masking the effects of the behavior estimation from the layered HMM.

As mentioned previously, the distance for which the future motion trajectory is sampled from is the fastest speed of the vehicle multiplied by 3 s. As vehicles often travel below the assumed fastest speed, the estimated risk values beyond 3 s remain reasonable, which explains the rising risk values even beyond 3 s before collision. It has also been



■ Fig. 57.16

Plot of mean and variance of risk estimation for Cross-Junctions. Source: (TAY 2009)

observed from the graphs that when vehicle V_A is going straight, the risk values beyond 3 s before collision rise slower and have greater variance compared to the other cases. It seems to suggest that under normal driving conditions, there is less of a chance of colliding with the vehicle in front. This can be explained by the fact that the model does not take into account behaviors such as sudden braking.

3.2 Driving Simulation

As it is difficult to perform experiments involving real life crash situations, experiments were performed in a virtual environment. The virtual environment is a virtual driving environment/simulator developed by Toyota Motors Europe.

The virtual environment is a geometrical model of the world in three dimensions, consisting of a road network populated with vehicles. To increase the realism of this virtual environment, with respect to risk estimation, the vehicles populated in the scene are driven by a human. The experiments in the simulator are performed in collaboration with Toyota Motors Europe and ProBayes.

3.2.1 Experimental Setup

The virtual environment simulates traffic conditions consisting of a number of vehicles traveling along the road network. Each of the vehicles is driven by a human driver. Each human driver controls his virtual vehicle via a wheel joystick as if the human driver is in the driver's seat. Recording a large scenario with many vehicles driven simultaneously requires a large number of human drivers and wheel joysticks. The scenarios are generated in an iterative fashion where only a single human-driven vehicle is recorded at a time. In each iteration consisting of a single human-driven vehicle, the previously recorded human-driven vehicles are replayed. The entire scenario is generated by this process of iteratively “adding” human drivers into the scene. Because of the virtual environment, crashes can be easily and safely created.

In a scenario, the risk of a single designated vehicle, V_A , will be evaluated. The risk to be evaluated is the same as that of the Monte Carlo experiments of [Sect 3.1.2](#), [Eq. 57.39](#). In contrast to the Monte Carlo experiments, where the behaviors of all vehicles are known, no behaviors on vehicles are known. A layered HMM evaluates the behavior for every vehicle present in the scene except vehicle V_A . Different time horizons for the evaluation of risks were also performed.

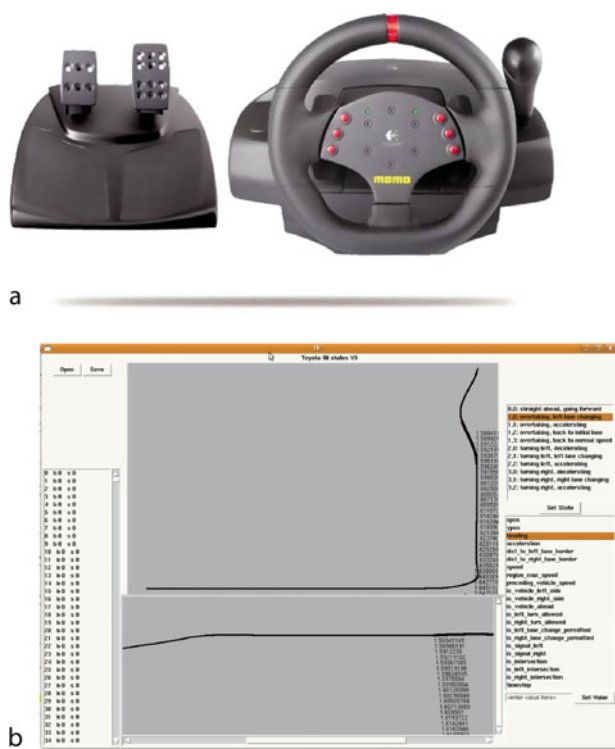
Training data was collected by collecting driving sequences from a number of human participants. Each participant uses the driving wheel ([Fig. 57.17a](#)) as an interface to the virtual environment to simulate driving from the point of view of the driver in 3 day. The set of collected driving sequences are then annotated manually using an annotation tool developed by ProBayes ([Fig. 57.17b](#)) before being used as the training data to train the parameters of the layered HMM.

3.2.2 Results

[Figure 57.18](#) shows the screen shot of the simulator. The simulator consists of a top-down 2D view of the environment, and a 3D view from the point of view of the driver. This 3D view window is also used by the human drivers when the human drivers are used to record the different scenarios. In the simulator, the risk calculated will be for the yellow vehicle whereas all other vehicles are red. For the experiments, the convention is that vehicles drive on the right lane.

In the 2D view, a color coded trail behind the yellow vehicle indicates the estimated levels of risk previously. The big yellow circle indicates the radius in which the red vehicles are taken into account. At all moments, the red vehicle nearest to the yellow vehicle will have its estimated layered HMM behavior probability displayed as vertical white bars. The 3D view indicates the speed of the yellow vehicle. The vertical color coded bar on the right gives the various risk value encoding from green being the least risky to red representing high risk. The vertical bar on the left indicates the current risk value for the yellow vehicle.

Current commercial crash warning systems are able to warn a driver if he is traveling too fast and about to collide with another vehicle in front. The risk estimation algorithm is

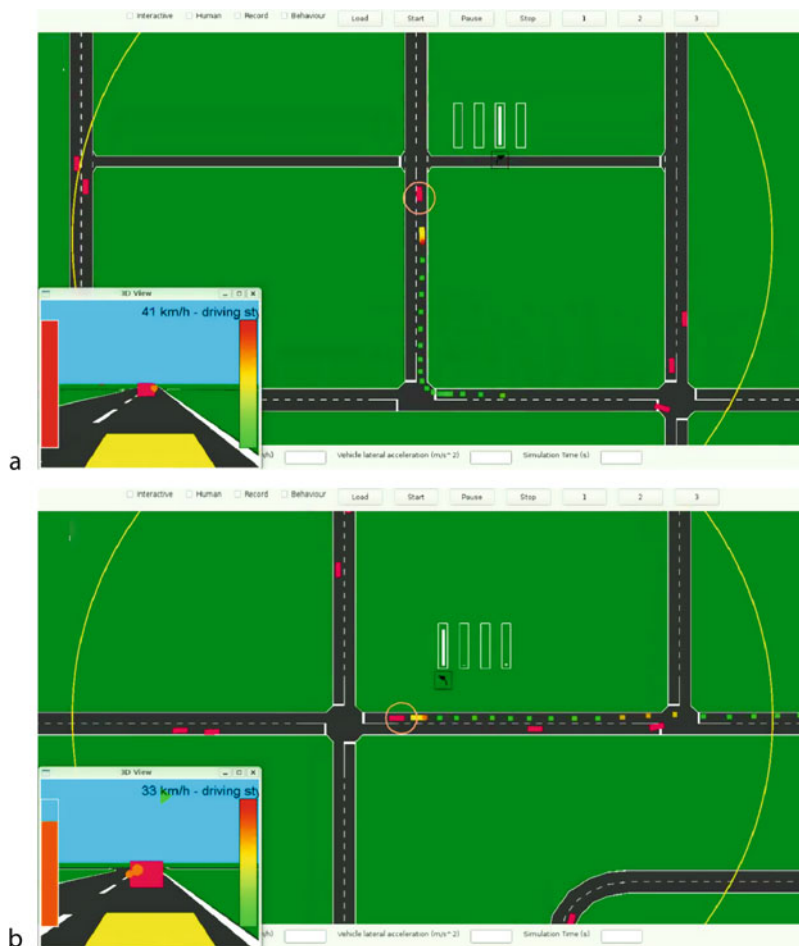


■ Fig. 57.17
Driving Wheel in (a) and Annotation tool in (b). Source (TAY 2009)



■ Fig. 57.18
Screen capture of simulator. Source: (TAY 2009)

able to provide the same functionality. ➤ *Figure 57.19* shows two such examples. In these examples, the red vehicles in front are estimated by the layered HMM to be turning left or right. The yellow vehicle has a relatively high speed with respect to the red vehicles and a high risk level is estimated as observed by the red vertical bar in the 3D view.

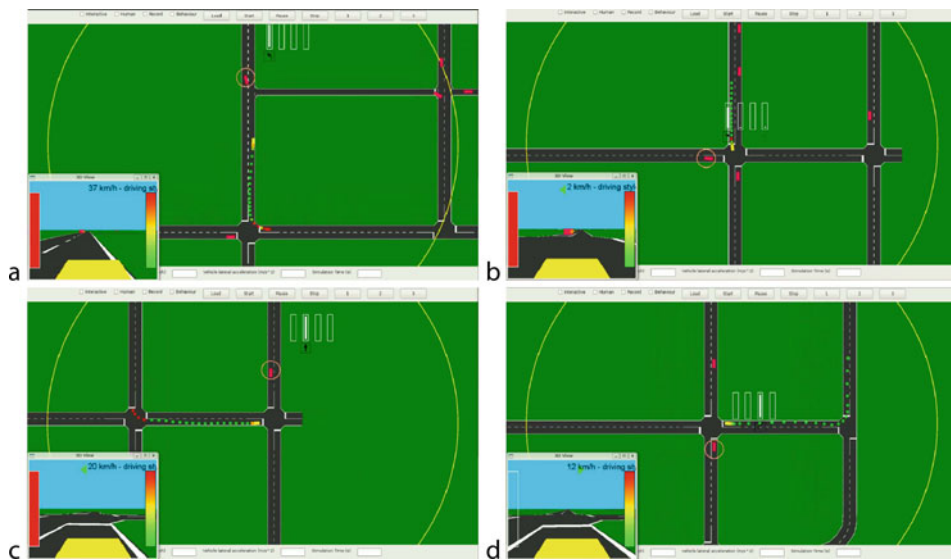


■ Fig. 57.19

Collision risks with vehicles in front. Similar to current state-of-the-art systems. Source: (TAY 2009)

The risk estimation algorithm is able to reason in more complicated situations such as the intersection. ➤ *Figure 57.20a, c* shows examples where an assumption on linear motion would not give a reasonable risk estimate when there are high risks of collision in reality. In these two situations, the layered HMM is able to correctly determine the behavior of the red vehicles. The combination of the behavior estimation from the layered HMM and taking into account the semantics (turning or going straight) at the geometrical level gives the appropriate high risk values.

➤ *Figure 57.20b, d* shows very similar situation at the cross intersection where the yellow vehicle and its nearest red vehicle might enter into collision. Again, by appropriately taking into account behavior probability distribution and geometry,



■ Fig. 57.20

Taking into account intersection when evaluating risk. Source: (TAY 2009)

reasonable risk probabilities are obtained. 🔍 *Figure 57.20b* presents a situation with a high risk of collision. In fact, regardless of the behavior of the red vehicle, chances of it going into collision are high. The situation for 20d looks very similar but it does not go into collision because the layered HMM had correctly recognized a turning right behavior for the red vehicle and does not enter into collision with the yellow vehicle. Current methods assuming linear motion will probably trigger a high false alarm in this case.

The application of Gaussian Process in modeling deviation from the center of the lane gives intuitive risk values. Vehicles which deviate far from the center of the lane pose non-negligible risks to other vehicles traveling on the adjacent lane. This effect can be seen from the sequence of 🔍 *Figs. 57.21a–c*. In general, the risk values for vehicles about to pass by each other are higher as the vehicles passing by each other get closer. As noted in 🔍 *Sect. 2.5*, the geometric configuration of the vehicle is taken into account. If geometry is not considered for example, a simple threshold on the euclidean distance between two vehicles consistently gives false alarms of high collision probability each time a vehicle is about to go past adjacent to another. 🔍 *Figure 57.21d* shows another instance where an assumption on linear motion will give a high probability of collision which is inaccurate. By adapting the Gaussian Process to the curved lane geometry of the road, a more reasonable risk value is obtained as can be observed by the left vertical bar on the 3D view.

🔍 *Figure 57.22* gives a quick summary on the recognition performance of the layered HMM as a confusion matrix. The confusion matrix is a visualization tool where the

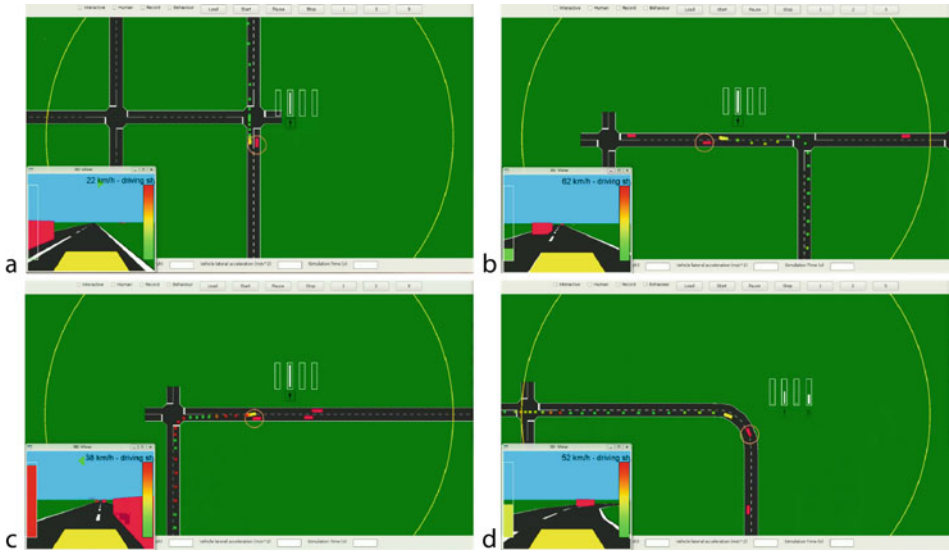


Fig. 57.21 Risks associated with passing by another vehicle along the adjacent lane. Source: (TAY 2009)

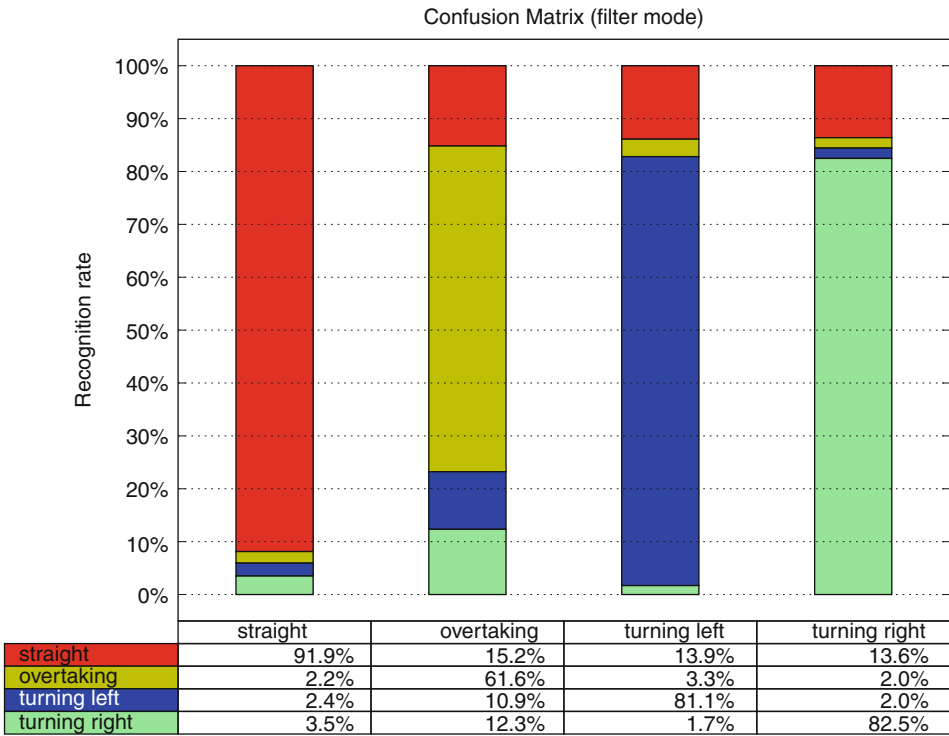
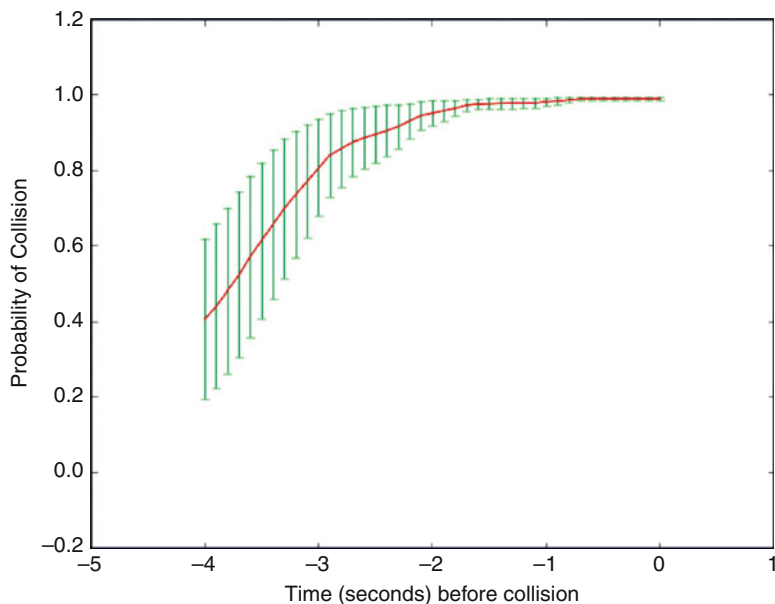


Fig. 57.22 Confusion matrix on the performance of the layered HMM. Source: (TAY 2009)



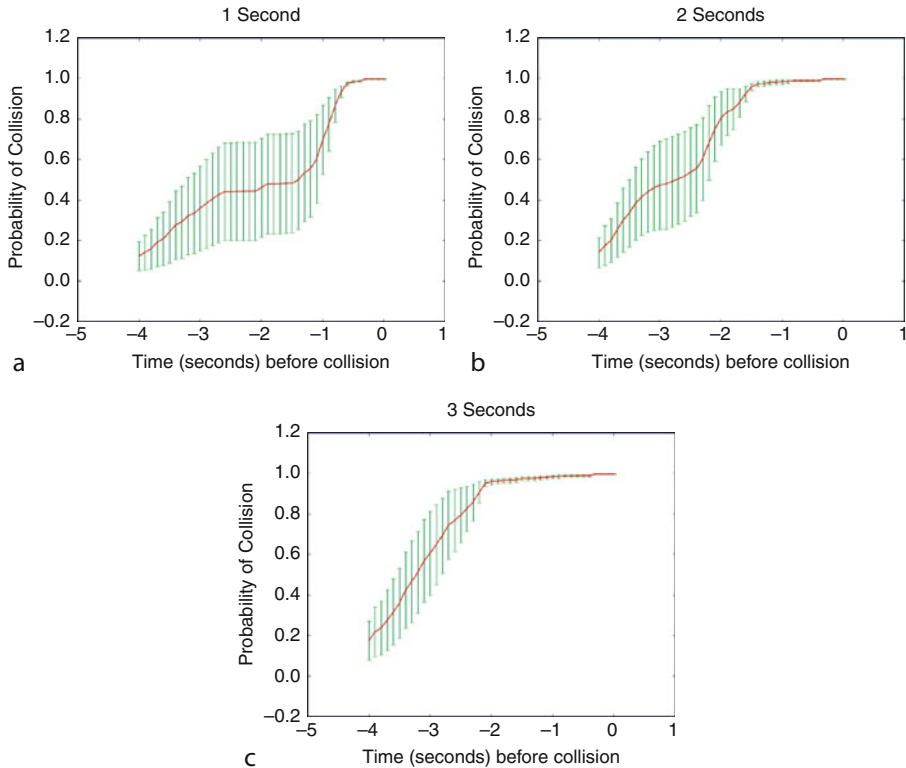
■ Fig. 57.23

Aggregate risk mean and variance for 10 human-driven scenarios. The time horizon for collision is 3 s. Source: (TAY 2009)

columns represent the true class and the rows represent the predicted class. From the confusion matrix, it is easy to see the percentage of mislabeling for each class and the diagonal of the confusion matrix represents the correctly predicted class. The highest recognition rate was for the going straight behavior, followed by the turning left/right behavior. The overtaking behavior had a relatively low recognition rate (61.6%). Intuitively, this is because it is easier to confuse the overtaking behavior which consists of lower level behaviors, lane changing, accelerating, lane change back to original, and resuming normal speed. These lower level behaviors can easily be mixed up with the other behaviors.

► [Figure 57.23](#) shows the plot of the risk values 4 s before each collision. The means and variances were computed using the vector of risk values 4 s before each collision across 10 different scenarios. The time horizon of the risk evaluation for the 10 scenarios was fixed at 3 s.

To evaluate the effects of different time horizons in the evaluation of risk, the experiments for the 10 scenarios were performed for different time horizons, for 1, 2, and 3 s. The means and variances 4 s before each collision across 10 scenarios are displayed in ► [Fig. 57.24](#).



■ Fig. 57.24

Plot of mean and variance of risk estimation for different estimation time horizons. Source: (TAY 2009)

4 Conclusion

In this chapter, the various modules involved in the estimation of risk in a structured road traffic situation are presented. A layered HMM (● Sect. 2.2) is used to estimate the behaviors of the vehicles in the scene. The behaviors can be separated into two hierarchical levels, corresponding to the architecture of the layered HMM. The high level behaviors are behaviors such as overtaking, turning left, etc. Each high level behavior is composed of a sequence of sub-behaviors.

For each of the higher level behavior, there is a corresponding Gaussian Process which is a Gaussian distribution over the paths of the typical pattern for each behavior (● Sect. 2.3). The Gaussian distribution over the future motion path (● Sect. 2.3.1) is obtained by first transforming the observations to a canonical space in which the *canonical GP* resides. The transformation is conformal and uses a discretized least squares approach to approximate the conformal transform in the form of a 2D mesh where the

transformation for each triangle mesh is approximately affine. The probability distribution over future motion is obtained in canonical space by conditioning the *canonical GP* on the transformations. The inverse conformal transformation is then applied to obtain the final probability distribution over future motion in world space for risk evaluation.

The risk is evaluated in a probabilistically sound manner (➊ Sect. 2.4), based on the Gaussian probability distribution over future motion for the various behaviors, and the estimated behaviors from the layered HMM. It has also been showed that with the combination of Gaussian Process and layered HMM to model motion at various semantic levels, there are many different ways of evaluating risk values each having its associated semantic, and is dependent on the application requiring the risk value.

References

- Brand M, Kettner V (2000) Discovery and segmentation of activities in video. *IEEE Trans Pattern Anal Mach Intell* 22(8):844–851
- Brand M, Oliver N, Pentland A (1997) Coupled hidden markov models for complex action recognition. MIT Media Lab, Cambridge, MA, pp 994–999
- Eck M, DeRose T, Duchamp T, Hoppe H, Lounsbery M, Stuetzle W (1995) Multiresolution analysis of arbitrary meshes. In *SIGGRAPH'95: Proceedings of the 22nd annual conference on computer graphics and interactive techniques*. ACM, New York, pp 173–182
- Fuerstenberg K, Chen J (2007) New european approach for intersection safety – results of the ec project intersafe. In: *Proceedings of the international forum on advanced microsystems for automotive application*. Springer, Berlin Heidelberg, New York, pp 61–74
- Galata A, Johnson N, Hogg D (2001) Learning variable length markov models of behavior. *Comput Vis Image Und* 81:398–413
- Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. *J Res Natl Bur Stand* 49:409–436
- Hongeng S, Brmond F, Nevatia R (2000) Representation and optimal recognition of human activities. In: *Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR 2000*, IEEE, Hilton Head Island SC, USA, pp 818–825
- Lee DN (1976) A theory of visual control of braking based on information about time-to-collision. *Perception* 5(4):437–459, Edinburgh
- Lvy B, Petitjean S, Ray N, Maillot J (2002) Least squares conformal maps for automatic texture atlas generation. In: *ACM SIGGRAPH proceedings of the 29th annual conference on computer graphics and interactive techniques*. ACM, New York, Jul 2002
- National Highway Traffic Safety Administration (2004) Vehicle-based countermeasures for signal and stop sign violation. Technical report DOT HS 809 716, NHTSA, U.S. DOT
- Oliver N, Horvitz E, Garg A (2002) Layered representations for human activity recognition. In: *Proceedings fourth ieee international conference on multimodal interfaces*, pp 3–8
- Pierowicz J, Jocoy E, Lloyd M, Bittner A, Pirson B (2000) Intersection collision avoidance using its countermeasures. Technical report DOT HS 809 171, NHTSA, U.S. DOT
- Pinkall U, Des Juni S, Polthier K (1993) Computing discrete minimal surfaces and their conjugates. *Exp Math* 2:15–36
- Rabiner L (1989) A tutorial on hmm and selected applications in speech recognition. *Proc IEEE* 77(2):257–286
- TAY C (2009) Analysis of dynamic scenes: application to driving assistance. PhD thesis, Institut National Polytechnique de Grenoble, France, Sept 2009
- Wilson AD, Bobick AF (1998) Recognition and interpretation of parametric gesture. In: *Sixth International Conference on Computer Vision*, IEEE, Bombay, India, pp 329–336, Jan 1998

Section 11

A Look to the Future of Intelligent Vehicles

Michael Parent

58 Legal Issues of Driver Assistance Systems and Autonomous Driving

Tom Michael Gasser

F4 -Co-operative Traffic and Driver Assistance Systems, Federal Highway Research Institute (BASt), Bergisch Gladbach, Germany

1	<i>Introduction</i>	1520
2	<i>Regulatory Law on Driver Conduct</i>	1521
2.1	Advanced Driver Assistance Systems (ADAS)	1523
2.2	Partially Autonomous Systems	1525
2.3	Highly Autonomous Vehicle Systems	1527
2.4	Fully Autonomous Vehicles	1528
3	<i>Liabilities</i>	1528
3.1	Road Traffic Liability	1528
3.2	Product Liability	1529
3.2.1	Advanced Driver Assistance Systems (ADAS)	1530
3.2.2	Partially Autonomous Systems	1530
3.2.3	Highly Autonomous Vehicle Systems and Autonomous Vehicles	1531
4	<i>Areas with Restricted Public Access</i>	1532
5	<i>Data Privacy</i>	1532
6	<i>Conclusion</i>	1534

Abstract: While legal issues of driver assistance systems appear to be largely solved for systems supporting the driver with information or such that remain easily overrideable/oversteerable, the increase in automation can eventually bring about a paradigmatic change of the “driving task”: Up to date, the driver’s responsibility for the use of systems is maintained, thus remaining within the traditional concept of driving. In future, however, a substantial further increase in automation can lead to a structural shift. The legal issues this change would raise must be identified and handled at an early stage of research to avoid false investment as well as inequitable legal consequences. The legal issues related to driver assistance and autonomous systems are thereby cross-sectional in nature and have a link to the issue of acceptance as far as a basic legal change is intended and necessary.

Cooperative systems, presently under intensive research with their plentiful possibilities and benefits, also give rise to legal uncertainties in several fields (including liability). However, here communication architecture, technical design, as well as potential operators vary tremendously depending on use case and information required for the respective function. This retards a universally valid description of accompanying legal issues. As long as these systems are, however, only meant to take effect by informing the driver on the oncoming traffic situation and dangers without suggesting full reliability (as is presently mostly under research), data privacy should prove to be the only important (but resolvable) issue.

1 Introduction

With the first Advanced Driver Assistance Systems (ADAS) intelligent technologies took effect in the sphere of the driver. Up to date ADAS support the driver, either by providing information or by directly supporting vehicle control. While before, up to the early 1990s, the task of driving was left completely to the skills of the driver, the ADAS emerging since have started participating in the driver’s role of accomplishing safe driving. This reasoning leads us to the necessity of looking into regulatory law on the conduct of the driver to answer the question in how far rules and regulations within regulatory law effect driver assistance systems. In future, the degree of automation these systems offer will increase and potentially enable systems to safely execute all the driver’s tasks under specified driving conditions without even requiring the driver to execute permanent control while operating. The driver would then – depending on the degree of automation – be left in a monitoring part.

Regulatory law on the conduct of the driver is therefore closely related to the field of system application. Most of all, the road traffic codes must therefore be investigated closely as to their potential impact on (the use of) intelligent vehicles. Furthermore the liabilities related to production and operation of such vehicles must be carefully assessed as this belongs to one of the most substantial risks intelligent vehicles face for implementation. Finally, a very specific point lies in the usage and processing of personal data. This is strongly related to cooperative technologies for driver information and not necessarily

an element of systems executing the primary driving task. Therefore the issue of data privacy will be discussed separately.

Today, traffic safety relies on technical safety as well as “safe” driving. Statistics thereby reveal that accidents today result in the very most cases from driving errors, whereas the technical defect has a low impact (Müller 2007). This characterizes the potential of intelligent vehicles in terms of accident reduction. Both, technical standards and rules as well as road traffic codes are oriented toward danger defense: While technical rules and regulations focus on the side of engineering, which today – among other things – have the objective to warrant a reliable transformation of a driver’s commands, the road traffic codes stipulate on the driver’s rights and duties in terms of safe driving. The driver today plays a central role in what is traditionally considered to compose road traffic, the safety of which therefore to greatest extent still lies in his responsibility.

2 Regulatory Law on Driver Conduct

Vital for the correct application of the legal requirements and regulations drafted below is to distinguish between traffic areas with restricted and unrestricted public access (whereby right of property alone is not decisive): Road traffic codes (as are licensing requirements) are limited to the grounds open to public traffic (Hentschel et al. 2011; Burmann et al. 2010). If public access is therefore excluded, these regulations will generally prove nonapplicable, e.g., on privately owned and fenced grounds where the owner can decide freely in own responsibility – on how to organize traffic.

In Europe, the USA, Japan, etc., the dangers arising from road traffic for the public are considered substantial and therefore governed in detail by the road traffic codes. These may, however, differ substantially between countries as is perhaps most obvious in the country-specific regulation for bidirectional traffic to keep either on the right- or the left-hand side of the road, respectively (so-called right/left-hand traffic countries). To allow for the minimum amount of unification necessary to maintain safe cross-border traffic, the law on the conduct of the driver is stipulated in international treaties requiring the governments to keep their road traffic codes in accordance with the rules internationally recorded.

The Vienna Convention on Road Traffic is an international treaty ruling *inter alia* on the regulatory law of driver conduct. Even though it has been negotiated in 1968 under the United Nations Economic Commission of Europe (UNECE), the Vienna Convention is a recent international treaty on road transport and ratified up to date (13th Feb. 2010) by 70 States worldwide (UNECE Status 2011). It does not take immediate legal effect for private bodies, but requires the ratifying countries to keep their rules and regulations (this includes road traffic codes regulating on drivers’ conduct) in accordance with the principles stated therein. In the interest of a broad international illustration in terms of the effect regulations on driver conduct take, the Vienna Convention on Road Traffic shall in the following be taken as a basis for the analysis of regulations suggested to be similarly in

place in most countries. Thereby, as already stated above, only national road traffic codes have immediate effect on the duties of drivers.

For the purpose of the following analysis on the effect of regulatory law on the conduct of the driver, the legal situation shall be contrasted with the effects intelligent vehicle systems take (intended to be used in traffic with unrestricted public access). In this context, it is the potential of intelligent vehicle systems to support or carry out vehicle control that is the subject matter of analysis.

Naturally, any legal rule of the past could not take the development on the way to intelligent vehicles into account as this must be considered unforeseeable at the time of negotiation/stipulation of traffic rules. Nevertheless the regulations must remain applicable in case of intelligent vehicles so as not to run into inconsistencies and contradictions hindering implementation and thus impairing benefits for traffic safety and thereby adding to the risks combined with the marketing of such vehicles.

The existence and role of the driver is most clearly stated in the following excerpt of the articles taken from the Vienna Convention:

- ▶ Art. 1 v) Vienna Convention on Road Traffic defines the driver:
 << [...] "Driver" means any person who drives a motor vehicle or other vehicle (including a cycle), or who guides cattle, singly or in herds, or flocks, or draught, pack or saddle animals on a road; [...]>>

- ▶ Art. 7 paragraph 1 reads:
 <<1. Road-users shall avoid any behavior likely to endanger or obstruct traffic, to endanger persons, or to cause damage to public or private property. [...]>>

- ▶ Art. 8 paragraph 1 and 5 read:
 <<1. Every moving vehicle or combination of vehicles shall have a driver.>>
 [...]
 <<5. Every driver shall at all times be able to control his vehicle or to guide his animals.>>

- ▶ Art. 13 paragraph 1 reads:
 <<1. Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all manoeuvres required of him. [...]>>

- ▶ Art. 14 paragraph 1 reads:
 <<1. Any driver wishing to perform an manoeuvre such as pulling out of or into a line of parked vehicles, moving over to the right or the left on the carriageway, or turning left or

right into another road or a property bordering on the road, shall first make sure that he can do so without risk of endangering other road-users travelling behind or ahead of him or about to pass him, having regard to their position, direction and speed. [...]>>

It can therefore be concluded that in terms of driver conduct it is the driver's obligation to remain in charge of his vehicle and override/oversteer inappropriate system interventions (Art. 8 and 13 Vienna Convention). This requires the availability of the option to override/oversteer in order to enable the driver to comply. On the other hand this assumes and requires the driver's permanent concentration in observing and reacting to the traffic situation he is confronted with. This element of permanent attention to the traffic situation, the lane or the intended driving corridor can also be deduced from Art. 7 paragraph 1 and Art. 14 paragraph 1 of the Vienna Convention.

In the past and insofar as ADAS reach, the demand of permanent control has led to the reasoning that systems must remain overrideable/oversteerable (Albrecht 2005). It has further been questioned whether non-overrideable/non-oversteerable systems can be considered permissible in the first place as the driver would not be in a position to comply with his regulatory obligations in terms of conduct. Naturally, this does not lie within the legal effect of a regulation tailored to regulate safe driver conduct. On the other hand, however, the required conduct might be made impossible by non-overrideable/non-oversteerable intelligent vehicle systems if such a design was chosen. Thus it can be considered a requirement possibly resulting from the principle of unity of the legal system not to override/oversteer the driver by intelligent vehicle systems as this might turn out contradictory to a driver's duties.

The second most important finding from the rules of conduct is the required permanent concentration of the driver on driving. This is a necessary precondition to be deduced from the very most rules on driver conduct. Of course, at the time road traffic codes have been recorded, autonomous driving did not have to be considered: It was in so much a matter of course that the driver was permanently needed to accomplish driving even over very short periods of time that there was no need whatsoever to consider even the possibility for autonomously controlled phases.

As far as driver assistance systems reach, these two principles deduced above from the regulatory law on driver conduct have not been put into question so far. As even the wording "assistance" clearly reveals, the driver is only supported by the technical means of ADAS. Up to date the driver remains in a superior position in relation to the system (i.e., the system permanently allows the driver to execute his will) and on the other hand does not allow the driver to draw his attention off the driving task.

2.1 Advanced Driver Assistance Systems (ADAS)

The definition of driver assistance is implicitly built upon the assumption that such a system permanently requires the driver in the loop either to accomplish the assisted task itself (in case of a purely informing system) or to fulfill another task in driving (e.g., if

longitudinal control is assisted by an Adaptive Cruise Control (ACC) System, lateral control remains to be fulfilled by the driver). Apart from this, it can be stated that in case of ADAS, system limits remain that require the driver to permanently monitor the amount of assistance provided and correct the systems' interventions. The support ADAS can make available is exercised in very close cooperation with the driver. Thereby, the system limits that require constant surveillance and correction can directly be traced back to the comparatively low degree of automation ADAS offer.

If these legal findings are analyzed on the level of advanced driver assistance systems available today, the validity of the above mentioned can be verified. Even though it can be questioned whether or not systems functioning on the stabilization level of driving can be considered to be ADAS in the first place, they nevertheless remain in accordance with the findings above: Even though the antilock braking system (known as "ABS") actually intervenes by reducing the brake pressure applied by the driver for split seconds, this takes effect by only optimizing the functional process of braking (and thus allows for shorter braking distances under certain conditions and, most of all, maintaining the possibility for the driver to steer). From a legal point of view, the support an antilock braking system therefore offers to the driver can be classified to only improve the functioning of the braking system (Albrecht 2005). For electronic stability control (ESC), conformity with regulatory law on driver conduct can be derived from the concrete design and function of the system: During normal driving, ESC continuously monitors steering and vehicle direction. The direction intended by the driver (determined through the steering wheel angle sensor) and the vehicle's actual direction (determined through lateral acceleration, rotation (yaw), and wheel speeds) are the basis for ESC interventions. An ESC will only intervene when the first sign for a loss of control over the vehicle is detected (i.e., when the vehicle is not going where the driver is steering). An ESC intervention thereby always remains in line with the will of the driver as it will only intervene according to the steering angle applied. Thus an ESC even enhances the driver's control over the vehicle up to the maximum of the physically defined limits the friction coefficient between tyre and road surface allows. What adds further to legal soundness of ESC interventions is that they only take place in emergency situations when the driver is at least about to lose control. Such a loss of control is in legal terms also to be presumed in the case of an emergency braking system: Here, in case of early interventions, the driver must still be given the possibility to override the system. However, as soon as the system can sufficiently detect that an accident has become otherwise unavoidable (e.g., by swerving) the intervention (by braking) is considered in line with regulatory law (from a product liability point of view there might, however, still be good reason to foresee a possibility to override/oversteer – but this is no longer a matter of regulatory law on drivers' conduct as there is no alternative to braking).

However, there have been suggestions in the past to equip vehicles with intelligent speed management systems that have been argued to be most effective in case of a non-overrideable/non-oversteerable design (e.g., non-overrideable/non-oversteerable intelligent speed adaptation). The pros and cons in terms of safety shall not be made an issue here. Yet, based on an assessment of regulatory law, it must be pointed out that such

a system design would interfere with the demand to keep the vehicle under full control of the driver: Choosing a non-overrideable/non-oversteerable design would at least limit the driver in his ultimate choice of speed and is therefore in conflict with regulatory law (Albrecht 2005).

In contrast, informing and intervening systems the driver can override/oversteer at any time while supporting in the primary driving task, remain in line with regulatory law on driver conduct: An example for such a system might be a lane change assistant supporting the driver with information on vehicles in the blind spot as well as on closing vehicles, possibly combined with an intervention in lateral guidance in case the driver yet pulls over toward an occupied lane. If looking closely at such a system, it turns out to apply a combination of two assisting strategies: The one lies in the information for the driver that a vehicle is either in the blind spot or approaching from behind in the next lane. The second is the correcting intervention, helping the driver to avoid a collision or any substantial interference with oncoming traffic. In the first case, it is naturally left to the driver to take up the information and decide upon how to accomplish his immanent primary driving task. Therefore, no conflict with full control over the vehicle is insofar identifiable. Furthermore, there is no conflict with the driver's duty to permanently survey lane and surrounding traffic: The system in no way hinders the driver to check himself whether or not the lane he intends to change to is clear or not (and the instructing manual will even require just this from him due to the supporting character of ADAS). The system insofar only provides for additional safety in case the driver (negligently) overlooks dangers and thus supports the driver in reducing his potential for driving errors. The same is true for the correcting intervention in case a collision or substantial interference with oncoming traffic is imminent: The driver can override/oversteer the intervention if he wishes (or needs) to do so (e.g., in order to avoid a more severe front-collision) so that the vehicle remains under his (superior) control. Furthermore, the intervention only assists in case the driver has already missed a vehicle – e.g., in the blind spot – so that the intervention supports in preventing the driving error already present to take further effect.

2.2 Partially Autonomous Systems

In the following the attempt shall be undertaken to describe the legal effect of a gradual rise in the degree of automation. Thereby, in contrast to ADAS already available on the market today, any details of autonomous systems remain subject to presumption until implemented. Therefore, in the following the autonomous systems and vehicles are in each case described in detail in order to reveal the technical design underlying each legal evaluation.

In comparison to ADAS, systems allowing for partially autonomous driving can be imagined that are characterized by a larger degree of automation. Furthermore, these systems can be imagined to combine longitudinal as well as lateral control simultaneously (Weilkes et al. 2005). From a technical point of view, the system limits can be presumed to decrease furthermore. Therefore, permanent execution of driver's control is no longer

necessary to run the vehicle. However, the system limits are still great enough to require the driver to permanently monitor the surrounding traffic situation and immediately intervene in case the system fails to handle a certain situation in traffic. In order to do so, the system would remain overrideable/oversteerable at any time.

From a regulatory law point of view, such a degree of automation will still remain very close to the legal situation for ADAS: The need for permanent monitoring of the system at its limits as well as the surveillance of the surrounding traffic and the lane in front is still very much in line with the duties assigned to the driver by the regulatory law on drivers' conduct. Furthermore, the requirement of full control on the side of the driver is not impaired as the possibility to override/oversteer is on hand.

It must, however, be assumed that a system with such a degree of automation – especially in terms of lateral control – is meant to bring along the benefit of being able to drive hands-off. Insofar, the drivers' control over the vehicle must be considered to be dependent on the respective traffic situation. Therefore, in case of potentially dangerous traffic situations (and due to the system limits) regulatory law on drivers' conduct will require driving hands-on in spite of sound system operation. In every other traffic situation, care must be taken that the drivers' capability to accomplish the driving task is not substantially delayed by driving hands-off. In addition, national road traffic codes might not allow for hands-off driving in the first place which must be carefully assessed (and cannot be described for each country).

Further dangers can be imagined in case partially autonomous systems would be misused beyond their system limits (e.g., by over-relying on the system/neglecting driver duties). Such behavior would be opposed to regulatory law on drivers' conduct and must therefore be considered contrary to duty (if not even an administrative offence). Thus this misuse of a system would already be implicitly interdicted as a lack of control must then be assumed on the side of the driver (which would be in conflict with Art. 8 paragraph 5 or Art. 13 paragraph 1 Vienna Convention – regulations suggested to be in place similarly in national road traffic codes).

In the past, platooning of heavy goods vehicles has played a major role in the discussion on a possible area of application for partially autonomous systems. Insofar it must, however, carefully be distinguished between the different degrees of automation involved: As long as a distance to the vehicle in front is maintained that remains safely manageable for the driver, the system must indeed be classified as (only) partially autonomous. However, most scenarios of platooning foresee substantial benefits in decreasing headways for the reasons of fuel economy, reducing the need for (scarce) space in dense traffic, etc. This necessarily leads to a loss in control on the side of the driver, in case human reaction times, under the condition of driving with only a few meters of headway, turn out to be shorter than required in order to override/oversteer the system (e.g., in case of an application of maximum brake power by the vehicle in front the possibility to swerve in order to avoid the accident is thereby cut off). Therefore, the requirement of driver control is no longer fulfilled under these circumstances (which would even belong to the normal operating mode). Furthermore, most road traffic codes must be presumed to explicitly require headways substantially greater than those

necessary to achieve the aforementioned benefits (cp. Art. 13 paragraph 3 Vienna Convention: “The driver of a vehicle moving behind another vehicle shall keep at a sufficient distance from that other vehicle to avoid collision if the vehicle in front should suddenly slow down or stop.”). This will in the very most cases result not only in a breach of drivers’ duties but will qualify as an administrative offence according to national law as well. To conclude on the actual degree of automation in the case of platooning, with a decrease in headway, it must be stated that such operation of a vehicle requires a higher degree of automation: The driver is inevitably cut off from the possibility to immediately intervene and handle the traffic situation from that point in which human reaction times turn out insufficient to override/oversteer (Frenz and Casimir 2009). This would suggest categorizing such systems as “highly autonomous,” however, much depends on the concrete design chosen in case of platooning.

To conclude on partially autonomous systems it can be established that this degree of automation largely remains in line with regulatory law on drivers’ conduct. Thereby, the situation for hands-free driving must be assessed according to national road traffic codes, but as long as technical design of the system allows the driver to fulfill his duties according to regulatory law on drivers’ conduct, no difficulties are to be observed in this respect.

2.3 Highly Autonomous Vehicle Systems

Highly autonomous vehicle systems again involve an even higher degree of automation. These systems might have a range in functionality that can be characterized by the absence of system limits during autonomous phases. They would not require the driver at all to immediately take over control, but would necessitate a takeover only after a “sufficient time buffer.” Meanwhile (i.e., during the automated phase) the driver would no longer be required to monitor neither the highly autonomous system, nor lane and surrounding traffic. The highly autonomous system would remain fully overrideable/oversteerable by the driver whenever he wishes to do so.

In case of this scenario, the full range of drivers’ duties is covered by the autonomous system. Furthermore the driver is not hindered in executing any driving maneuver as he can permanently override/oversteer. However, the purpose of the system lies in relieving the driver from permanently controlling and surveying the system as well as lane and surrounding traffic. Yet, neglecting permanent surveillance of lane and surrounding traffic – according to present regulatory law on driver conduct – must be considered a breach of the driver’s duties.

Apart from this, the question may arise, whether such highly autonomous systems would in terms of system control during highly autonomous phases remain bound by regulatory law on the conduct of the driver. In contrast to the aforementioned degrees of partially autonomous systems and ADAS, the scenario of highly automated vehicles at hand does not intend permanent control of the driver in the first place. So in case running such a system would be permissible without control of the driver (which would be the case if the regulatory law ruling on drivers’ duties would be amended in order to allow for the

full potential of such a system), it will turn out that in fact there is no regulatory law on how highly autonomous system control is to be applied. Only factual necessities of mixed autonomous and conventional driving would force to adjust system control to the rules stated in the regulatory law of driver conduct (i.e., the road traffic codes).

2.4 Fully Autonomous Vehicles

Insofar as fully autonomous vehicles, i.e., the full automation of vehicles can be considered a reasonably foreseeable scenario at all, the degree of automation would thereby reach an upper level. In this case no circumstances would be imaginable that require a takeover by the driver for safety reasons. It must even be presumed that such a degree of automation would cope with defects and other unforeseeable events by returning to the “safe” condition – i.e., the standstill of a parking vehicle. Yet, it seems absolutely feasible to allow for the permanent possibility of the driver to override/oversteer.

In terms of regulatory law on the conduct of the driver, the findings are therefore identical with the autonomously controlled phases in case of highly automated vehicle systems: The only purpose of such a system would lie in the possibility to relieve the driver from his task of driving. Nevertheless, the duties in terms of present regulatory law would persist so that there would be the need for a legal revision to realize the full benefits such systems offer.

The design of system control would in case of fully autonomous vehicles likewise be bound to regulatory law on driver conduct only by factual necessities. Such factual necessities might turn out to be mixed traffic of autonomously and conventionally controlled vehicles.

3 Liabilities

In order to fully assess the situation in terms of foreseeable liability issues of intelligent vehicles, road traffic liabilities must be covered just as well as product liability.

3.1 Road Traffic Liability

In most European Countries a regime of strict liability – i.e. liability regardless of fault – has been installed. These liabilities are, however, still regulated on a national state level within Europe and have only been harmonized in terms of basic principles to ensure for a reliable compensation in case of accidents. Therefore, in absence of a common basis for road traffic liability, the legal situation cannot be analyzed with universal validity in the following brief outline. The situation valid for Germany shall therefore only highlight possible contradictions of highly autonomous systems and fully automated vehicles which might similarly be encountered elsewhere. The detailed legal analysis in terms of road traffic liability must therefore be dealt with on a national level in each country.

German law rules the liability of the “Halter.” The “Halter” is defined as the person who makes use of the vehicle at his own expense, i.e., the person who decides upon the use of the vehicle (as a source of hazard) which allows the allocation of liability to the “Halter” as a result of risk initiation. He is strictly liable for the operational risk of the vehicle, no matter, if faulty behavior of the driver – who might, but need not at the same time be “Halter” – or a technical defect of the vehicle itself has led to the accident. The risk of liability assigned to the “Halter” therefore covers the complete operational risk (Hentschel et al. 2011; Burmann et al. 2010). As far as this strict liability for the operational risk is concerned, any kind of driver assistance system as well as any kind of automation would be covered by the liability of the “Halter” according to the present legal situation.

In order to make this strict liability tolerable for the “Halter,” the third party insurance is mandatory. The insurance is – according to German law – even jointly and severally liable toward any third party in every case of a damage.

As far as foreseeable at present, this situation can be transferred to any kind of ADAS as well as automation without discrepancy.

The driver is also liable according to German law, but only according to law of negligence. Therein – and this is remarkable when it comes to automation – negligence of the driver is assumed whenever it comes to an accident. The burden of proof is therefore on the driver to prove absence of negligence (Hentschel et al. 2011; Burmann et al. 2010). Here the question arises, whether this can be regarded appropriate in cases of the aforementioned degrees of highly autonomous vehicle systems as well as fully autonomous vehicles. In these cases the drivers’ attention would be dispensable during automated driving phases. Therefore, it would be contradictory to assume negligence of the driver in case an accident occurs during such automated phases. Yet, this would be presumed according to the present legal situation. A corrective might be achieved by technical means, e.g., accident data recording (with a downside in terms of data privacy as well as the “*nemo tenetur se ipsum accusare*” – principle in criminal law establishing that there is no legal obligation to self-incrimination). Otherwise the driver is endangered to carry the burden of proof for not having acted negligently (what he might, however, in (far) future be able to accomplish by the *prima facie* evidence of an autonomous system being installed in the vehicle he has been driving).

As a result, only the – possibly unique – presumption by German law that the driver has acted with fault in case of an occurring damage would be contradictory. The strict liability of the “Halter” would remain applicable no matter which degree in automation is reached.

3.2 Product Liability

Product liability can also differ between the Member States of the European Union; however, this field of liability has been subject to harmonization according to EU-directive 85/374/EWG. Therefore, a regime of strict liability has been installed in all countries within Europe. The main legal requirements are thereby the defectiveness of the product

as well as damage causation by the defect. Possible defects leading to the assumption of defectiveness result – as far as relevant here – from faulty system construction as well as insufficient instruction of the product user.

3.2.1 Advanced Driver Assistance Systems (ADAS)

Product liability naturally applies to ADAS as to any other consumer product. ADAS thereby remain limited in performance due to their aforementioned low degree in automation. Here it must be kept in mind that driver assistance systems are only meant to support the driver. ADAS are furthermore characterized by system limits that must be taken into account by the user as characteristic behavior to be considered and adapted to (according to the instructions in the product manual). Therefore (and if the system remains permanently overrideable/oversteerable), a compensatory reaction of the driver can be expected whenever necessary (without considering this system performance to establish a defect).

This necessitates that the intervention remains “controllable” for the driver which is directly related to the capability of the (human) driver to handle critical situations. This concept of “controllability” has therefore been made a central principle of ADAS design within the “Code of Practice for the design and evaluation of ADAS.” This document is originally the result of the horizontal activity “Response 3” within PREVENT, an integrated European project. Its aim has been to make risks of product liability manageable for manufacturers when developing and testing ADAS (RESPONSE 2009). One of the main guiding principles therein for the design of ADAS is to assess whether the driver is able to handle critical situations in case an ADAS reaches its limits. If this is the case, the scope of the legal element “defectiveness” is thus reduced to mainly those cases beyond driver’s control.

3.2.2 Partially Autonomous Systems

As already stated above, partially autonomous systems are characterized by a larger degree of automation. They might combine longitudinal as well as lateral control simultaneously and there are less system limits. Even though the driver does not have to execute permanent control to run the vehicle, he is still required to permanently monitor the surrounding situation in traffic and to override/oversteer if the system fails to handle a traffic situation properly.

From a product liability point of view, the element of permanent monitoring of surrounding traffic and system operation by the driver is decisive. The partially autonomous system thereby remains very close to the situation valid for ADAS insofar as the driver is expected to override/oversteer any inappropriate system operation. The significant difference of the partially autonomous system toward ADAS lies in the fact that the driver is no longer taking action to operate the vehicle (e.g., by steering, braking, and accelerating)

once the partially autonomous system has been activated. The main difference therefore arises from the fact that the driver is taken out of the loop of active driving and only required to monitor system operation – thereby obliged to intervene immediately in case system operation seems inadequate for the immanent situation in traffic.

The main risk thereby lies in the fact that the driver might thus be enabled to turn to other tasks than driving or even unintentionally drop in his attention toward the monitoring task (which would no longer – as is the case with conventional driving – be accompanied by an immediate effect on vehicle guidance and would therefore potentially remain unnoticed by the driver). For this reason, the above-mentioned principle of controllability is substantially complicated in case there is a need for the driver to take over control. Furthermore, this leads to the question – to be answered beyond the legal scope – whether an immediate takeover of control can sufficiently be expected from the driver by overriding/oversteering. Finally, the technical possibilities to ensure driver attentiveness and thus keeping the driver in the loop of permanent system monitoring (which must remain in line with data privacy issues) cannot be foreseen yet. The further research on these questions concerning situation awareness and human abilities in monitoring system operation as well as technical means of preventing negative effects of driver inattention to carry into effect are vital for the further legal assessment of product liability risks.

The legal issue of product liability affected by partially autonomous systems is the question of what will therefore qualify as “reasonably foreseeable misuse” of the product user (i.e., the driver) and therefore lies within the sole responsibility of the manufacturer (in terms of product liability and only in case it comes to a damage). This must be distinguished from what must be considered “abuse” of a partially autonomous system which would in contrast lie in the responsibility of the user (i.e., the driver) and does not lead to product liability of the manufacturer (again only if it comes to a damage). Therefore there remains substantial need for technical consolidation and further development in order to fully assess the legal risks in terms of product liability for partially autonomous systems.

3.2.3 Highly Autonomous Vehicle Systems and Autonomous Vehicles

Highly autonomous vehicle systems or even autonomous vehicles involve a yet higher degree of automation. Technically these systems are characterized by the absence of system limits during autonomous phases. Thereby, highly autonomous vehicle systems require a takeover by the driver only after a “sufficient time buffer”, whereas autonomous vehicles would not necessitate a driver takeover for safety reasons at all (these systems would presumably cope with unforeseeable events by returning to the “safe” condition – i.e., the standstill of the parking vehicle). Both degrees of automation, this is decisive in legal terms, would no longer require the driver to monitor neither the autonomous system or vehicle nor lane and surrounding traffic.

Therefore, a very high level in automation is presupposed and necessary to make these systems or vehicles feasible for publicly accessible road traffic. Thereby the highly autonomous systems would still necessitate a takeover by the driver at the limits of the system. A “sufficient time buffer” would then be available to the driver in order to adapt to the traffic situation he has not been monitoring during the preceding automated phase. Again, therein lies an issue in terms of driver controllability. This controllability is – different from the one present with partially autonomous systems – most dependent on the driver’s ability to recover in terms of situational awareness necessary for driving.

Autonomous vehicles on the other hand no longer require any driver takeover. For this reason the requirements in terms of technical reliability are tremendously high. At the same time, the risk of misuse and abuse by the driver in executing vehicle control is eliminated.

From a product liability point of view, the autonomously controlled phases without any need for monitoring by the driver put high demands on the reliability of system control. Damages occurring during these phases that are no longer controllable by the driver, would – from what can be foreseen according to the present state of law – render the system/vehicle defective in terms of product liability.

4 Areas with Restricted Public Access

Quite in contrast to the legal issues outlined above stands the legal situation in areas with restricted (non-)public access. Here road traffic codes and vehicle licensing requirements do not take effect. As described above, the main idea behind the principles in regulatory law as well as technical requirements lies in danger-prevention that is implicated by multiple use of the road by the public. As long as public access is restricted, these legal obligations will usually prove to be nonapplicable.

In this case (exclusion of public access), safety can be achieved under the sole responsibility of the owner, operator, or carrier. Thereby, e.g., segregation of automated traffic from public access, the detection of trespassers, or the elimination of trespassing altogether can be considered appropriate technical concepts to implement the duty of care required for automation. The adherence to safety rules thereby remains vital as the present state of the art in science and technology must be implemented. Otherwise liability risks on the side of the operator or carrier will arise.

5 Data Privacy

Data privacy is a legal requirement resulting from human rights guaranteeing self-determination and freedom of the individual. The importance is substantially increasing in the world of information technology. The increase of intelligence in all kinds of vehicles makes it necessary to circumscribe basic legal elements that must be considered before developing any kind of system – otherwise running the risk of encountering rejection on the side of potential users as well as legal inadmissibility of certain strategies and

technologies that might be considered for implementation in an intelligent vehicle (e.g., as might be the case with certain designs of driver surveillance). On the other hand, it shall be stressed that data privacy is a principle of utmost importance in system design but cannot be considered any kind of limit in system permissibility: It must rather be understood as to be a design concept that must be properly implemented and considered during system development. The legal issue of data privacy is most often encountered in the discussions on cooperative (“C2X,” car-to-car and car-to-infrastructure) technologies. It is, however, not of greater importance here, but simply encountered frequently.

For intelligent vehicles, data privacy takes effect very much according to the nature of data processed and further depends on the full particulars of data processing foreseen. As far as the intelligent vehicle systems already dealt with in the chapter “Legal Issues” (i.e., ADAS, partially autonomous systems, highly autonomous vehicle systems and fully autonomous vehicles) are concerned, the particulars necessary for a specific legal evaluation are presently not foreseeable (autonomous systems might even take absolutely no effect in terms of data privacy at all). The same applies to cooperative system concepts allowing for the generation and transmission of information from vehicles to others, road side infrastructure, traffic management centers, and vice versa: Here the discussions on which information shall be transmitted, the necessary operator models, etc., have not yet reached a level of maturity that would allow for a substantiated legal analysis (the same applies to the other legal issues of cooperative systems, i.e., liability and the intended use of these systems in traffic). Therefore, in the following, only an outline on the basic principles of data privacy as far as relevant for the design of intelligent vehicle technologies is intended.

Within Europe, the minimum standard of data privacy (“data protection”) is stipulated by EU Directive 95/46/EG. This directive has been issued in 1995 to ensure a basic level of data privacy in the processing of personal data throughout Europe. In case of electronic telecommunications, the directive on privacy and electronic communications (2002/58/EG) must be considered preferentially as the superordinate directive in this specific field. This shall be taken as a basis for the following general outline, as a more detailed description must be refrained from at this point of technical research and development.

The application of data-protection law is limited by the range of personal freedom as “data protection” is based on basic human rights. Therefore the directive is – in spite of the misleading expression “data protection” – limited to the protection of data related to the individual. Generally, the processing of data is only permissible for conclusively enumerated reasons and beyond that legitimate only in case of consent by the holder of rights. Therefore, as a first conclusion, only personal data lies within the scope of the data protection directive and any data processing that is not ultimately necessary should be refrained from (principle of data avoidance/data economy). Furthermore, data can only be processed for the reason already underlying its collection, subsequent changes of the originally intended purpose at the time of data collection is cut off (limitation to purpose). The possibility to gain an informed consent from the holder of rights will in general legalize data processing. This, however, has its downside in user acceptance and practicability.

Bearing in mind these limitations in terms of data protection for the development of intelligent vehicle systems will already take substantial influence on the legitimacy of processing personal data. However, the final evaluation necessary to identify compliance must be left to a detailed analysis on the system level, taking the precisely defined personal data as well as purpose of acquisition and use into account. This must then be made subject to an evaluation according to national data protection rules that will apply in the single case.

6 Conclusion

First of all it must be distinguished between autonomous driving in public places (open to public access) and areas of restricted (nonpublic) access. The legal relevance of issues in regulatory law is given only for areas accessible to the public.

Furthermore, from a legal point of view on research and development, relevant issues for autonomous systems must be systemized according to their level of automation. Legally relevant conflicts in terms of regulatory law on driver conduct that are no longer manageable with present legal instruments occur in case of systems intended to automate such that the driver may turn away from his driving task. The mere possibility to do so is, however, already relevant in terms of product liability. And it must be pointed out that high degrees of automation require a very high level of functional safety as the driver is not available on short term to correct or intervene in upcoming situations which are excluded by system limits. This leads to an increase in the product liability risk on the side of manufacturers. The situation for autonomous driving in case of liabilities within road traffic must be analyzed on a national level in each country as legal regulations can differ immensely in this respect.

The legal situation for cooperative systems is mostly limited to the effects of data privacy as long as the systems only take effect by informing or warning the driver without arousing the (usually wrong) appearance of full reliability. According to present data privacy legislation, this issue remains resolvable but should be taken into account in system design at an early stage of research and development.

References

- | | |
|--|--|
| <p>Albrecht F (2005) Die rechtlichen Rahmenbedingungen bei der Implementierung von Fahrerassistenzsystemen zur Geschwindigkeitsbeeinflussung. Deutsches AutoRecht (DAR), pp 186–198</p> <p>Burmam M, Heß R, Jahnke J, Janker H (2010) Straßenverkehrsrecht Kommentar. C. H. Beck, München. ISBN 978-3-406-59421-2</p> <p>Frenz W, Casimir-van den Broek E (2009) Völkerrechtliche Zulässigkeit von</p> | <p>Fahrerassistenzsystemen. Neue Zeitschrift für Verkehrsrecht (NZV), pp 529–534</p> <p>Hentschel P, König P, Dauer P (2011) Straßenverkehrsrecht Kommentar. C. H. Beck, München. ISBN 978-3-406-60991-6</p> <p>Müller T (2007) Fahrerassistenz auf dem Weg zur automatisierten Fahrzeugführung. Automobiltechnische Zeitschrift (ATZ), pp 58–64</p> <p>Integrated Project PREVENT, horizontal activity “RESPONSE 3,” updated version 5.0 (2009)</p> |
|--|--|

Code of practice for the design and evaluation of ADAS. http://www.acea.be/index.php/files/code_of_practice_for_the_design_and_evaluation_of_adas/. Accessed 12 Feb 2011

United Nations Treaty Collection. (<http://treaties.un.org/Home.aspx>). Status of Convention on Road Traffic. [http://treaties.un.org/Pages/](http://treaties.un.org/Pages/ViewDetailsIII.aspx?&src=TREATY&mtdsg_no=XI~B~19&chapter=11&Temp=mtdsg3&lang=en)

[ViewDetailsIII.aspx?&src=TREATY&mtdsg_no=XI~B~19&chapter=11&Temp=mtdsg3&lang=en](http://treaties.un.org/Pages/ViewDetailsIII.aspx?&src=TREATY&mtdsg_no=XI~B~19&chapter=11&Temp=mtdsg3&lang=en).

Accessed 13 Feb 2011

Weilkes M, Bürkle L, Rentschler T, Scherl M (2005) Zukünftige Fahrzeugführungsassistentz – Kombinierte Längs- und Querregelung. Automatisierungstechnik (at) 53(1):4–10

59 Intelligent Vehicle Potential and Benefits

Claude Lurgeau

Mines ParisTech, Paris, France

1	<i>Fuel and Energy Problems</i>	1538
2	<i>Pollution and Greenhouse Gas Effect</i>	1538
3	<i>Safety</i>	1539
4	<i>Congestion: A Better Use of Infrastructure</i>	1540
4.1	Lateral Control for Augmented Productivity of Road and Parking	1541
4.2	Longitudinal Control to Improve the Traffic Throughput	1541
5	<i>Mobility for All</i>	1544
5.1	Multimodality	1545
5.2	Carpooling and Car Sharing	1545
5.3	Personal Rapid Transit (PRT)	1547
5.4	Demand Responsive Transit (DRT)	1549
5.5	Aging/Handicapped Society	1550
6	<i>Conclusion</i>	1551

Abstract: In this chapter, we will examine how the development of ITS can help solve major problems of humanity which are energy, climate change, congestion, and safety.

The coming decades will accompany a mobility revolution and will see the emergence of a new relationship of citizens with their automobile.

The concept of sustainable mobility covers deployment of electro-mobility but also new forms of relationships between citizens and nature.

Information and communication technologies will help promote multimodality, ticketing, car sharing, carpooling, PRTs, and progressive automation of driving through the dissemination of Advanced Driver Assistances Systems (ADAS) and cooperative systems (V2V and V2I).

1 Fuel and Energy Problems

All fossil fuels (coal, oil, gas) are the result of the transformation of plants or animals, after millions of years. However, in less than a century, mankind has almost exhausted these precious resources. When we say that it is humanity – this is not true. On the blue planet Earth, there are a little over seven billion people and about a billion vehicles, but they are not evenly distributed among populations. Rich countries that is to say North America, Western Europe, and Eastern Asia have more than 600 million vehicles (60% of total world stock), while their population is less than 1 billion, about 15% of world population.

To bring China to the same level of equipment than rich countries, it will require the equivalent of the total stock of vehicles that now exist on earth.

The production capacity of the top 20 world automakers is about 70 million cars per year. This can be considered very low, since that makes only one new car produced per year per 100 inhabitants, but on an other side, that is too much when you know that we will not have enough raw materials to produce them, and no fuel to make them roll.

In rich countries, the ratio of vehicles per inhabitant is about 55–60 vehicles per 100 people. These vehicles do not work more than half an hour per day in average. This means that they represent a bad investment for the owners. A new type of relation between the human and the vehicle has to be proposed. What is useful is not the vehicle as an object but the mobility service that it provides to the human.

The information and communication technologies cannot solve the problem of energy but can optimize the consumption particularly taking into account digital map and communication between vehicles and infrastructure.

2 Pollution and Greenhouse Gas Effect

Ten or fifteen years ago, experts all agreed on the fact that the key problems to solve in ground transportation systems were: safety, congestion, energy, and pollution. The perception of the problems by the society has changed; these four problems are still the same but the order has changed and now we would say: energy, environment, mobility, and safety.

Before the industrial revolution and the automobile era, animals and vegetal live in harmony on the earth.

Animals are ambulant chemical reactors which burns their foods with oxygen they breathe and reject carbon dioxide. In a complementary way, vegetal species breathe carbon dioxide to fix carbon and reject oxygen.

During the last century, mankind has accumulated aggression on nature. Humans have proceeded to a colossal deforestation to develop intensive agriculture, and consequently lower the potential of oxygen production.

In parallel, humans have burned millions of tons of petrol and rejected billions of cubic meters of carbon dioxide in the atmosphere, creating the greenhouse effect phenomenon.

Let us make a simple estimation of CO₂ expelled in the atmosphere per year by the billion of vehicle. We find a thickness of 1.6 mm of CO₂ all around the earth that justify plainly the expression greenhouse!

World car fleet	1,000,000,000
Average distance in kilometer covered by a car per year	10,000
Average CO ₂ emission in gram per kilometer	100
Total mass of CO ₂ in gram	1,000,000,000,000,000
Volume of CO ₂ in cubic meter	509,090,909,090
Surface of the earth in square meter	314,160,000,000,000
Thickness of CO ₂ in millimeter	1.62

Another way to challenge our imagination is to express by its weight the CO₂ released by a car. A very good modern car, which emits only 150 g of CO₂/km, makes a block of ice of 15 k, clinging behind your car after a run of 100 km.

There is no doubt that the CO₂ emission of one billion vehicles year after year modifies the climate.

3 Safety

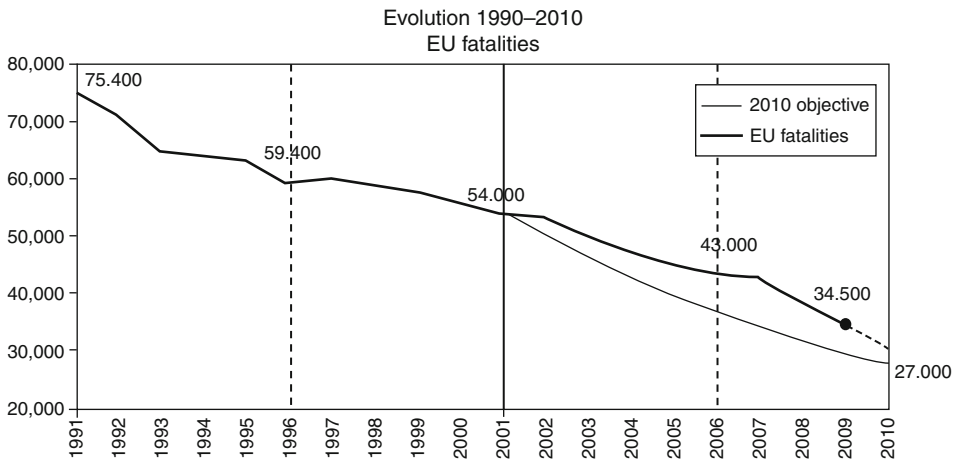
Regarding safety, the world can be divided in two categories:

- Advanced countries that is North America, Western Europe, and Eastern Asia which have the safest results
- The emerging countries which have still unsafe traffic results

In USA, there were 32,788 road deaths in 2010, a 3% drop from 2009. Traffic fatalities fell to an all-time low in 2010 even as Americans drove more miles.

Concerning Europe, we observe a drastic diminution of road fatalities while the traffic increases continuously.

Eastern Asia that is Japan, South Korea, and Australia represents about 10,000 fatalities per year.



■ Fig. 59.1
Evolution of fatalities in European Union

This means that all the rich countries together have about 100,000 fatalities per year. The total world fatalities are about 1,200,000 deaths per year.

This means that rich countries which have about 60% of the total number of vehicles, have less than 10% of the fatalities.

The reasons are clearly the quality of roads network, the quality of the vehicles, and education level of drivers.

When an accident occurs, there are four main causes: the vehicle, the infrastructure, the driver, and other cause. We know that in all countries, at any time more than 97% of causes of accidents are due to the driver. So a direction of progress is clearly to reduce progressively the role of the human in driving.

4 Congestion: A Better Use of Infrastructure

When a car moves, it occupies a safety bubble, which is rectangular in shape. The width of the bubble is equal to the width of the roadway on which the car moves. The length of the bubble depends on the speed of the vehicle. The approximate length of the bubble is the product of the speed of the vehicle by the driver's reaction time. This length is the safety distance. The consequence is that the floor is a critical resource that needs to be optimized.

In human driving, it can reach a throughput of 2,200 vehicles/h at a speed of 55 km/h.

When the car does not roll, it does no service to its owner. However, it occupies a parking space which is a considerable cost in cities.

Congestion is caused by an imbalance between supply and demand for space required for circulation in a particular place at a given time. To solve that imbalance, we can use two methods.

- The first one is to increase the traffic surface. This is what we have always done for one century, but the roads and highways are expensive and disfigure the landscape.
- The second is to reduce demand by taxing the right to drive and the right to park.

The intelligent car allows a third solution by increasing the productivity of existing infrastructure through the mastery of lateral control and longitudinal control.

4.1 Lateral Control for Augmented Productivity of Road and Parking

The highways and roads are made of lanes whose width is standardized. This width is 3.5 m in Europe and 4 m in America.

But the average width of the cars is approximately 1.75 m. This means that one can practically roll two cars in front on the same lane.

At least, on a two lanes road it is possible to drive three cars instead of two.

This bad utilization of the road is linked to the limits of human being who is unable to drive even at low speed of 60 km/h, a car in a tunnel where there would be 25 cm of free space on each side of the vehicle. A robot can do that without difficulty.

Thus, the intelligent car that has a precise lateral control can increase the productivity of existing infrastructure by about 50%.

The construction of roads represents substantial economic costs. A gain of 50% in usable area represents a considerable saving for the community.

In the same way, when the car is stationary, it occupies a parking space. In assuming that the car is driven by a human operator, it must at least increase the width of the car door open to allow the driver to leave his vehicle. If the conduct is automated, passengers can get off the vehicle and reduces the width required for parking.

4.2 Longitudinal Control to Improve the Traffic Throughput

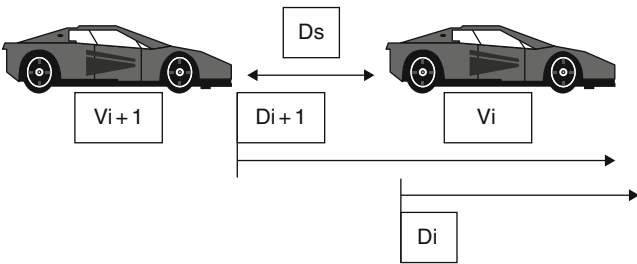
We have seen that each moving vehicle has to be surrounded by a bubble of safety. The wideness of the bubble can be lowered thanks to lateral control and that improve the productivity of the road by a minimum factor 1.5.

We can observe a higher benefit with longitudinal control.

Let us consider two vehicles V_i and $V_i + 1$ which move at the same speed V . If the leading vehicle V_i brakes at its maximum braking capacity to stop, the vehicle $V_i + 1$ which follows it must be able to stop without hitting it. The vehicle speed V_i decreases according to a linear law: $V(t) = V - \gamma_i \cdot t$ and it stops after a time $T = V/\gamma_i$, after traveling a distance

$$D_i = \frac{V^2}{2\gamma_i}$$

where γ_i is the deceleration braking of vehicle i (● Fig. 59.2).



■ Fig. 59.2
Computation of safety distance

Given the reaction time of the human driver τ , the vehicle $(i + 1)$ has already covered $V \tau$ before starting to brake. Then he stops by the same process as the vehicle i , that is to say, he travels a total distance:

$$D_{i+1} = V\tau + \frac{V^2}{2\gamma_{(i+1)}}$$

To avoid the impact, it is therefore necessary to have a safety distance D_s such that:

$$D_s > V\tau + \frac{V^2}{2} \left(\frac{1}{\gamma_{(i+1)}} - \frac{1}{\gamma_i} \right)$$

The minimum safety distance is the sum of two terms:

- A term related directly to the human operator and the limits of his reflex time. The parameter τ is estimated between 1 and 1.5 s. It is this term that can be lowered significantly in the event of an automated driving. In the calculation of safe distances, the administration takes an upper bound of 2 s.
- A term that expresses the differential braking capabilities of the two vehicles. This term may be negative if the leading vehicle has a braking capacity more efficient than the following vehicle or positive otherwise. In the first case, the theoretical safety distance must be increased in the second, it can be lowered. The problem is that the follower driver does not know whether his brakes are more or less efficient than those of his predecessor. To comply with the precautionary principle, it should be placed in the most adverse conditions.

If we assume that the vehicles have exactly the same braking capabilities, then the safety distance is limited to the first term alone $D_s = V \tau$.

Taking $\tau = 2$ s and if we express V in Km/h, then $D_s = 20/36 * V$. We obtain safety distances to different legal speed limits.

Speed in kilometers per hour	50	70	90	110	130
Safety distance in meter	28	39	50	62	73

These are the safe distances that we are obliged to comply under the new rules.

The drivers know how fast they roll through the tachometer, but they have no instrument to give them their safety distances except for luxury vehicles equipped with adaptive cruise control. The respect of safety distance is a matter of appreciation of the driver.

In rain or wet conditions, the braking capacities are lowered and the safety distances have to be increased. Generally, the public authorities lowered the speed limits by at least 20 km/h.

For example, on motorways the speed limit is lowered from 130 to 110 when it rains.

In the safety distance equation, if we devise each member by the speed, we obtain the minimum time T_{\min} between two vehicles.

$$T_{\min} = \tau + \frac{V}{2} \left(\frac{1}{\gamma_{(i+1)}} - \frac{1}{\gamma_i} \right)$$

The worst situation is obtained if the leading vehicle has an infinite capacity of braking.

This means that if it stopped instantaneously, the vehicle behind must stop before a collision occurs. This is called a “brick-wall stop.”

$$T_{\min} = \tau + \frac{V}{2} \frac{1}{\gamma_{(i+1)}}$$

This hypothesis maximizes the safety but is not realistic since the passengers of leading vehicle will die as a consequence of infinite deceleration.

If we take a braking capacity of 0.5 g that is about 5 m/s² then $T = 2 + V/10$ in seconds.

Speed in kilometers per hour	50	70	90	110	130
Safety time in seconds	3	4	5	5	6
Safety distance in meter	47	77	113	154	203

The safety distance with the brick wall hypothesis increases in a considerable manner and the throughput of the road becomes unacceptable. So we will assume that all the vehicles have the same braking capacity or that each of them knows the breaking capacity of the others – this will become possible in the future with vehicle to vehicle communication.

The tip-to-tip headway expressed in seconds is simply defined as:

$$T = \frac{L}{V} + \tau + \frac{V^2}{2} \left(\frac{1}{\gamma_{(i+1)}} - \frac{1}{\gamma_i} \right)$$

where L is the average length of a vehicle.

The vehicular capacity of a single lane of vehicles is simply the inverse of the tip-to-tip headway. This is most often expressed in vehicles per hour that is :

$$N_{veh} = \frac{3600}{T}$$

Neglecting the term L/V , the vehicular capacity of a lane is approximately

$$N_{veh} = \frac{3600}{\tau}$$

That means the limitation of the capacity is related to the human driver reaction time. With $\tau = 2$ s, we obtain 1,800 vehicles per hour and per lane but if we use automated longitudinal control with a τ which can be for example 500 ms, then the capacity becomes 7,200 vehicles per hour and per lane. We have increased in a significant way the productivity of the infrastructure.

5 Mobility for All

In human history, men have always needed to move from a place to another one. At the beginning, they had only their legs to move, then they have domesticated animals such as horse, camel, elephant, ass to move faster and with less fatigue and also to transport heaviest burdens.

The discovery of steam engine during nineteenth century allowed the dissemination of railways in Europe first then all over the world. This has been a fabulous adventure, since men have been able to move in 1 day what necessitates 1 week before. But trains had and still have two major constraints which limit the human satisfaction. These two constraints are spatial and temporal, spatial since the train does not always stop where you live and where you go, and temporal since it does not operate the time you wish.

The automobile overcomes these two constraints. The road network can be seen like a graph in which vertices are parking lots or garages and edges are segment of road. A car owner can leave his garage at any time to reach another parking lot anywhere in the network by different routes. So the automobile offers to its owner a total freedom and the pleasure to drive. That explains the phenomenal development of automotive industry all over the twentieth century.

- But nobody had imagined the problems of energy, environment, safety and congestion which have risen. And now we have to find solution to these problems and imagine new solutions of mobility.

Mobility for all does not mean a vehicle for every citizen. A vehicle represents between several months and several years of work of a human being, and generally a car is used less than half an hour per day – so on the economical point of view it is a poor investment at the individual level. It is also a bad investment for the society.

Nevertheless, the need of mobility must be satisfied and ITS can contribute in an efficient way to that.

First, we have to notice that the urbanization increases continuously. The ratio of urban to rural way of life is about 50 to 50 now and it will reach 60 to 40 in 2050. The most efficient solution to satisfy mobility in dense urban area is public transportation by trains, metro, tramways. In big cities where there is a good public transportation network, many people give up owning a car.

But in suburban and rural areas, the car is often the only solution to reach a good quality of mobility service.

5.1 Multimodality

Multimodality is an important concept which will benefit of information and communication technologies. A multimodal door-to-door journey combines different forms of transport, taking into account traffic congestion, environmental impact, cost, time, comfort, and accessibility (► Fig. 59.3).

For passengers transport, multimodality solution embraces all type of mobility such as buses, taxis, train metro, walking, cycling ... A significant impediment to multimodal travel is the fragmentation of information about the various resources, their availability, their costs, and the ease of interconnectivity.

Reliable cross-mode timetables and seamless ticketing are essential to maximize consumer acceptance of public transport as a viable alternative to personal vehicle ownership.

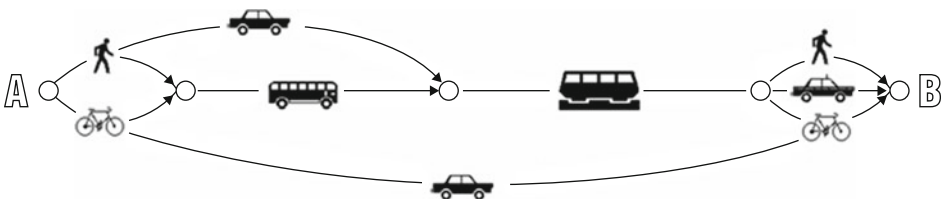
5.2 Carpooling and Car Sharing

http://en.wikipedia.org/wiki/Car_sharing

These two concepts although different have the same goal which is to maximize the productivity of a vehicle seen as a critical and rare resource.

In carpooling, we maximize the number of passengers in a ride, while in car sharing we use the vehicle as long as possible to improve its productivity. Obviously, we can fuse the two concepts to obtain the best use of car.

Car sharing is a model of car rental where people rent cars for short periods of time, often by the hour. They are attractive to customers who make only occasional use of a vehicle, as well as others who would like occasional access to a vehicle of a different type than they use day-to-day. The organization renting the cars may be a commercial business or the users may be organized as a democratically controlled company, a public agency, or



■ Fig. 59.3
Illustration of multimodality concept



■ Fig. 59.4

VELIB: Velib is a form of car sharing service for bicycles in Paris since 2007. Velib is compatible with the Navigo pass of Paris Metro. The city of Paris is organising AUTOLIB on the same concept with electric vehicles

a cooperative, ad hoc grouping. Today, there are more than 1,000 cities in the world where people can share a car (► Fig. 59.4).




Car sharing differs from traditional car rentals in the following ways:

- Car sharing is not limited by office hours.
- Reservation, pickup, and return are all self-service.
- Vehicles can be rented by the minute, by the hour, as well as by the day.
- Users are members and have been pre-approved to drive (background driving checks have been performed and a payment mechanism has been established).
- Vehicle locations are distributed throughout the service area, and often located for access by public transport.
- Insurance and fuel costs are included in the rates.

Carpooling is a simple solution which is not technical but is common sense. It is a mode of travel in which several users share a given journey on one vehicle rather than riding alone and separately in their own cars.

Hitchhiking is an ancient form of carpooling. Unlike the hitchhiking approach in which the hitchhiker does not participate to fee, carpooling is planning to travel between those who negotiate cost sharing and decide who provides the vehicle and the driver.

To measure the effectiveness of the solution, let us take a simple example in which 40 people living in the same residential area move to the same area of economic activity every working day. Without public transportation, the most effective solution would be to charter a bus of 40 seats. The worst solution is that 40 people go to work alone in their car. Carpooling is the solution in which people travel with 10 cars.

	Usual solution
	The best solution
	Car pooling solution

■ Fig. 59.5

Illustration of carpooling concept

It is observed that car sharing provides positive answers to all problems:

- It has declined by 75% in energy consumption.
- It has declined by 75% emissions of greenhouse gases.
- It has divided by 4 the number of cars in the traffic flow.
- It has reduced the likelihood of accidents.

At the individual level, 40 users have divided their transportation costs by 4 (➤ Fig. 59.5).

There are several variants of carpooling:

- Carpooling which corresponds to regular displacements from home to work place
- Carpooling which corresponds to occasional but longer distance trip
- Dynamic carpooling in which, both the driver and the hitchhiker connect in real time

The Internet is a powerful tool to promote the many variants of carpooling.

Specialized web sites, establish the connections between vehicle owners making long journeys, and potential hitchhikers.

This type of service has a real hit with young people who can move with limited budgets.

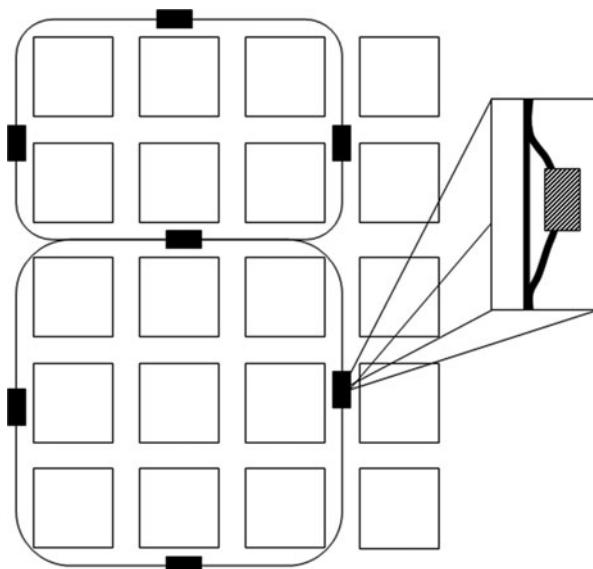
The owner announces its vehicle type, the route, schedules, and costs that can be shared by many. The “Carpoolers” must register on the site, but registration may be free. The site maintenance is funded by advertising and is a showcase of know how to sell private sites to companies or communities.

A financial incentive to develop carpooling is the creation of carpool lanes on toll roads. The HOV (“High Occupancy Vehicles”), lanes are widespread in North America where urban toll highways are numerous.

5.3 Personal Rapid Transit (PRT)

http://en.wikipedia.org/wiki/Personal_rapid_transit

Personal rapid transit or pod car is a public transportation mode featuring small automated vehicles operating on a network of specially built guide ways. PRT is a type of



■ Fig. 59.6

Personal rapid transit seen as a bidimensional horizontal lift

automated guide way transit, a class of system which also includes larger vehicles all the way to small subway systems.

In PRT designs, vehicles are sized for individual or small group travel, typically carrying no more than 3–6 passengers per vehicle. Guide ways are arranged in a network topology, with all stations located on sidings, and with frequent merge/diverge points. This approach allows for nonstop, point-to-point travel, bypassing all intermediate stations. The point-to-point service has been compared to a taxi or a horizontal lift (elevator) (► Fig. 59.6).

The main characteristics that define a PRT are:

- Vehicles are fully automated, that is to say without a human driver.
- The vehicles run in dedicated site, on a mesh of lanes.
- The vehicles are small size (2–20 people).
- The service schedule can be planned and fixed to the peak hours or on demand at off-peak hours.
- Users have to go up and down at fixed points like bus stops or taxi ranks.
- At breakpoints, vehicles have to leave the track so as not to prevent other cars to double.
- The electric power used reduces pollution and noise.
- PRT is completely automated in terms of driving, routing, and ticketing.
- The movement of vehicles is coordinated to optimize the flow.
- Vehicles may travel in packs, allowing the dynamic creation of trains (► Fig. 59.7).

A pilot system at London Heathrow Airport, United Kingdom, was constructed using the Ultra design.



ULTRA station at Heathrow



2 getthere in MASDAR

Fig. 59.7

Two examples of PRT : ULTRA at Heathrow and 2 getthere in Masdar

A PRT system (by 2getthere) went into operation in Masdar City in the UAE in November 2010. The system has ten passengers and three freight vehicles serving two passengers and three freight stations connected by 1.2 km of one-way track. The system is in operation 18 h a day, 7 days a week serving the Masdar Institute of Science and Technology. Trips take about two and half minutes (i.e., an average speed of roughly 12 mph/20 km/h) and are presently free of charge. Average wait times are expected to be about 30 s.

5.4 Demand Responsive Transit (DRT)

(http://en.wikipedia.org/wiki/Demand_responsive_transport)

DRT is a user-oriented form of public transport characterized by flexible routing and scheduling of small/medium vehicles operating in shared-ride mode between pickup and drop-off locations according to passengers needs.

DRT systems provide a public transport service in rural areas or areas of low passenger demand, where a regular bus service may not be as viable.

A DRT service will be restricted to a defined operating zone, within which journeys must start and finish. Journeys may be completely free form, or accommodated onto skeleton routes and schedules, varied as required. As such, users will be given a specified pickup point and a time window for collection. Some DRT systems may have defined termini, at one or both ends of a route, such as an urban center, airport, or transport interchange for onward connections.

DRT systems require passengers to request a journey by booking with a central dispatcher who determines the journey options available given the users' location and destination.

DRT systems take advantage of fleet telematics technology in the form of vehicle location systems, scheduling and dispatching software, and hand-held/in vehicle computing.

Vehicles used for DRT services will generally be small minibuses, reflecting the low rider ship, but also allowing the service to provide as near a door-to-door service as practical, by being able to use residential streets.

For a model of a hypothetical large-scale demand-responsive public transport system for the Helsinki metropolitan area, simulation results published in 2005 demonstrated that “in an urban area with 1 million inhabitants, trip aggregation could reduce the health, environmental, and other detrimental impacts of car traffic typically by 50–70%, and if implemented could attract about half of the car passengers, and within a broad operational range would require no public subsidies.”

5.5 Aging/Handicapped Society

<http://www.tc.gc.ca/eng/innovation/tdc-summary-12900-12927e-920.htm>

Worldwide, the population is aging, and the segment of the population older than 80 years old is increasing fastest of all. Because disabilities increase with age, the demand for accessible transport is expected to grow. Mobility is important for daily living, but people increasingly will have to stop driving because of health problems.

Aging causes physiological changes that make driving more difficult. These include increased reaction time, deteriorating vision particularly at night, and a reduced ability to split attention between several tasks. Accident rate per mile increases as driver's age past about 65, and increases rapidly beyond about 75 years old.

ITSs have much to offer people with impairments. For drivers, ITSs can partly compensate for the physiological changes that make driving more difficult for older people while improving everyone's safety. The application of ITSs in public transport improves the efficiency of transit operations and enables the provision of multimodal trip planning information. Real-time information can be provided at bus stops and stations, in vehicles, and in the home (via the Internet and pagers).

Personal vehicles account for more than 80% of trips made by older people. For seniors, the use of a personal vehicle is the single most important factor in maintaining an independent way of life.

A new class of vehicle will be required to provide independent local mobility for those who can no longer drive an automobile.

The application of ITSs to guide visually impaired people as pedestrians and through terminals is under way. The requirements of elderly and disabled people must be incorporated during the development of ITS applications and in the presentation of electronic information.

6 Conclusion

One hundred years ago, in the early twentieth century, one could still find horses, carriages, stagecoaches and dung in major world cities.

In few decades, an entire economy, a whole way of living and traveling, has disappeared, giving way to the car.

As we enter the twenty-first century, the situation of car manufacturers is somewhat comparable to that of diligence manufacturers a century ago. An important revolution of mobility will take place in the coming decades.

In this chapter, we have shown how the deployment of intelligent transportation systems can help solve major problems of mankind, including energy, pollution, greenhouse gas emissions, urban congestion related to traffic, and road safety.

The dissemination of vehicles with low CO₂ emissions, and particularly electrical vehicles, will grow, but sustainable mobility will also be deployed in the form of multimodality, car sharing, carpooling, and solutions leveraging on information and communication technologies.

References

Laurgeau Claude : Le siècle de la voiture intelligente – Presse des Mines, Paris. ISBN:978-2-911256-10-3

Towards a European road safety area: policy orientations on road safety 2011–2020. http://ec.europa.eu/transport/road_safety/pdf/com_20072010_en.pdf

Headway. <http://en.wikipedia.org/wiki/Headway>

Car sharing and car pooling. http://en.wikipedia.org/wiki/Car_sharing

Personal rapid transit (PRT). http://en.wikipedia.org/wiki/Personal_rapid_transit

Demand responsive transport. http://en.wikipedia.org/wiki/Demand_responsive_transport

Aging/Handicapped society. <http://www.tc.gc.ca/eng/innovation/tdc-summary-12900-12927e-920.htm>

60 Applications and Market Outlook

Michel Parent

Unité de recherche INRIA Paris Rocquencourt, Project IMARA,
Le Chesnay Cedex, France

1	<i>Introduction</i>	1554
2	<i>Private Passenger Vehicles</i>	1555
3	<i>Public Urban Vehicles</i>	1562
4	<i>Urban Freight Delivery</i>	1564
5	<i>Rapid Transit (Including BRT, PRT, and CTS)</i>	1566
6	<i>Long-Distance Freight</i>	1572

Abstract: The technologies for implementing intelligence in road vehicles are improving very rapidly and products are now available in the market. Not only for traditional passenger vehicles but also for trucks and buses and for a new generation of urban vehicles that are now appearing on the market as a transportation service instead of a product to be bought by the final user.

This chapter will look into the potential markets for the technologies in these different applications such as:

- Private passenger vehicles
- Public urban vehicles
- Urban freight delivery
- Rapid transit (including BRT, PRT, and CTS)
- Long-distance freight

1 Introduction

Since the beginning of the automobile, the manufacturers have proposed constant improvements of their products on the market with two primary objectives (as with any other mass product): reduced manufacturing cost and improved performances.

The performances searched for the automobiles were initially focused on the ease of operation (hence the first major breakthrough was the electric starter) and the operating performances. Under this term we mean the top speed but also the comfort and the safety (meaning essentially braking and handling performances).

Lately, performances focused mainly on passive safety and efficiency (in terms of energy cost but more and more in terms of emissions because of stringent regulations) but still including comfort in terms of ride and ease of operation.

In this light, the arrival of new “intelligent functions,” allowed by the availability of powerful and low-cost electronics, is bringing a complete new line of market potentials for the automobile manufacturers and their suppliers.

However, in order to appreciate the market outlook of these new functions (which can also bring new products in terms of vehicle designs), we have to segment this market in very specific areas where the goals are quite different in terms of cost-benefit ratios. So, this chapter will look into the following markets:

- Private passenger vehicles
- Public urban vehicles
- Urban freight delivery
- Rapid transit (including BRT, PRT, and CTS)
- Long-distance freight

2 Private Passenger Vehicles

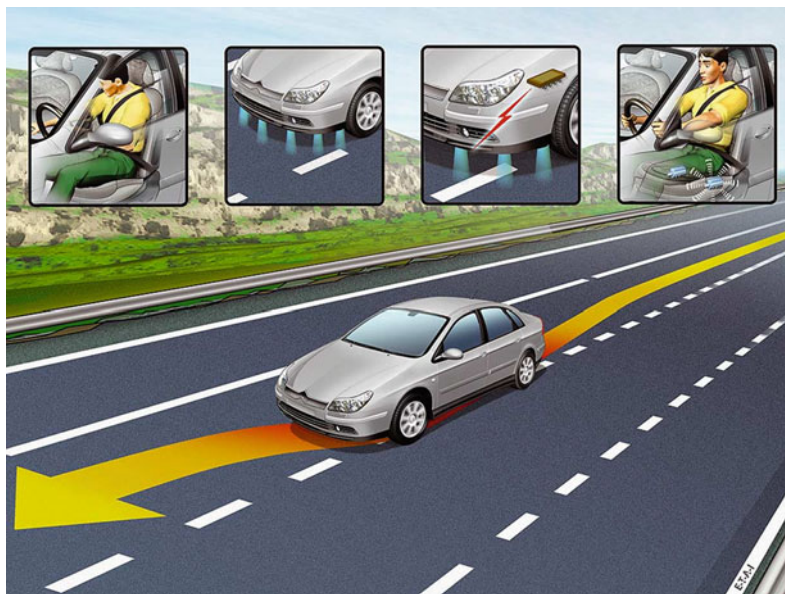
The twenty-first century has seen the arrival en masse of electronics in the general automotive industry. This has started previously with the introduction of antilock systems for the braking (often called ABS although this is a brand name from Bosch) and electronic injection that both started for high-end vehicles in the 1960s. In both cases, we can start to talk about “intelligence” since these systems process data gathered from sensors and act physically on the vehicle through a complex algorithm. This trend is now continuing with more and more complex systems such as electronic stability programs (ESP), intelligent automatic gearboxes (some of which can even anticipate the needs for shifting through a connection with the map data), adaptive suspensions, etc. However, all these systems are almost invisible to the user except through improved performances in terms of handling, braking, or power, often barely perceptible, except in case of emergency (e.g., when braking on slippery roads).

However, new electronic functions are now proposed to the car buyer as more explicit drivers’ assistance (or ADAS: Advanced Drivers Assistance Systems). The goals are still the same as before for this market segment: better safety, better comfort, or better performances. But, at the same time, this market segment does not allow to change radically the paradigm of the private automobile that must show a strong continuity in its design as well as in its operation. Therefore, the controls of the vehicle must remain sensibly the same (i.e., steering wheel and pedals but also all the accessory controls), both to satisfy the demand for continuity and sometimes for regulatory reasons (hence the impossibility at the moment to introduce “steer-by-wire”).

These new “intelligent functions” can be divided into three major classes depending on the level of assistance brought to the driver:

- Informative
- Assistive
- Autonomous

Informative systems. In the informative class we find the assistance brought to the driver by way of sensory information: visual, auditory, or haptic (vibrations on the seat or torques on the steering wheel or forces on the pedals). It is the responsibility of the driver to take action. A typical example is a navigation system which helps the driver to find his/her way in the best fashion (faster, nicer, cheaper, etc.). Other informative systems include warning systems about potential hazards such as lane departures or dangerous lane changes (blind spot detector) or excessive speed (by regulation or if a slower vehicle is ahead or if the road is slippery) or obstacle detection during maneuvers. All of these types of assistance depend mostly on information coming from localization and from digital maps and, obviously, a lot of processing but for warning systems, they also depend on sensory equipment such as cameras, radars, ultrasounds, and infrared sensors. Emergency



■ Fig. 60.1

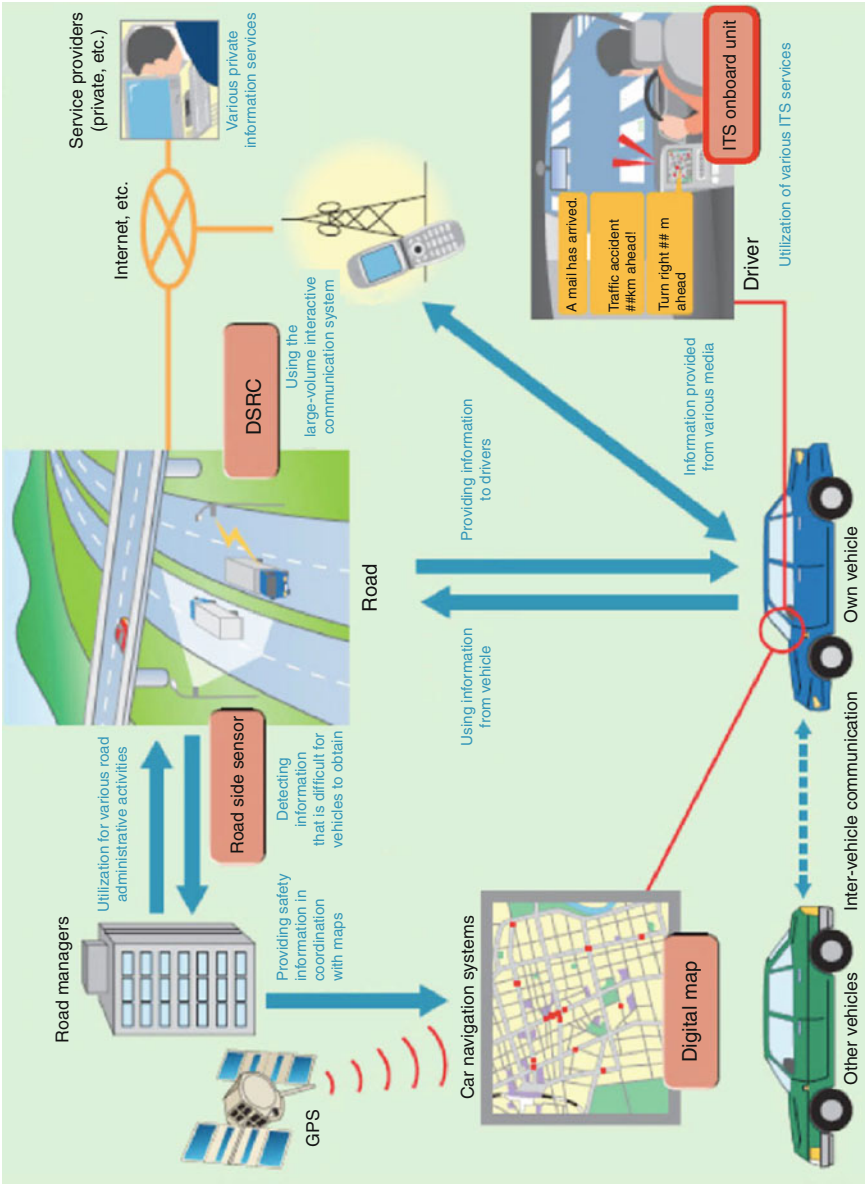
Lane departure warning on a Citroen with seat vibration (Citroen)

call systems (or eCall) that send an automatic call with the localization of the vehicle to an emergency center in case of an accident also fall under this category (● Fig. 60.1).

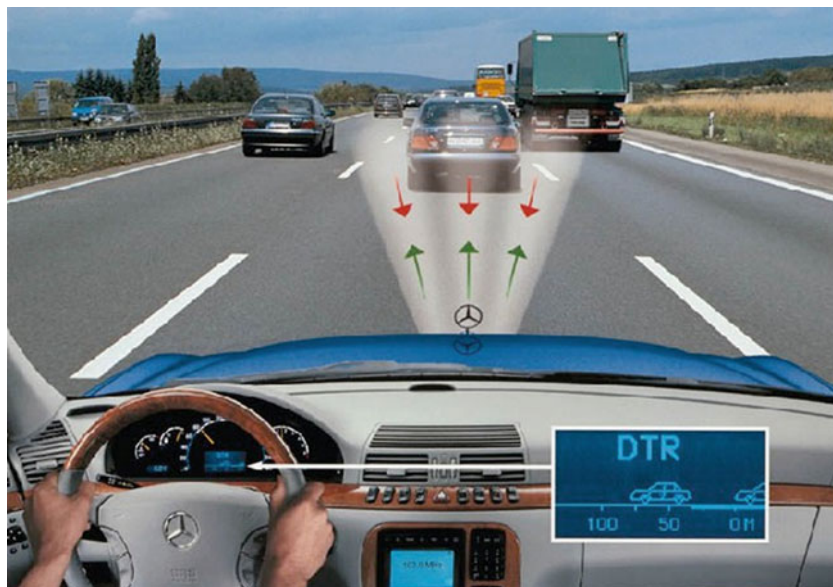
A new form of assistive system relies on the communication of real-time information from the infrastructure or, for future systems under test, from other vehicles. Through standardized data exchanges (see the CALM standard, Williams 2005), the vehicles can be aware of problems on the road ahead, such as accidents or sudden slowdowns or incoming vehicles or pedestrian at intersections. Each vehicle (which also acts as a probe vehicle to send data) can then compute a risk factor and warn the driver appropriately. Such systems are currently under deployment in Japan (Smartway) (● Fig. 60.2).

Assistive systems. In this class, we find truly drivers' assistance system in the sense that the system acts on the driving functions of the vehicle, whether it is on the steering, the acceleration, or the braking. However, this assistance is only temporary and all of the time (except for emergency situation) fully under the driver's control.

The first assistive system truly labeled an ADAS has been the so-called ACC (Adaptive Cruise Control), which acts on the acceleration and brakes to maintain a set speed unless there is another vehicle in the same lane at a lower speed in which case, it will maintain a set distance (or time gap). These ACC started to appear in Japan in the late 1990s with laser scanners to detect the position of vehicles ahead. Later, systems with radar sensors were introduced in Europe. Although the first systems performed well only in limited situations such as good visibility, no rain, straight highway, and high speed and limited deceleration, the new systems work rather well in a wider range of situations such as



■ Fig. 60.2
Smartway system in Japan (Smartway Project)



■ Fig. 60.3

ACC from Mercedes (Mercedes)

stop-and-go traffic and curvy roads and in different weather situations and can even handle emergency braking in case of unavoidable collision (collision mitigation) (🔗 Fig. 60.3).

Often found in complement to ACC is the lane-keeping function that tends to keep the vehicle in the middle of the lane defined by its road markings and detected by one or several cameras (prototypes also work with laser scanners). After processing the video image(s), the system applies a torque to the steering column that tends to put the vehicle back on its proper track. However, the driver has still full control and can override this torque. Actually, the driver **MUST** apply some torque at frequent intervals to demonstrate to the system that he (she) is still in control. This is to prevent the driver from falling asleep or to think that the car can drive by itself without supervision (🔗 Fig. 60.4).

A similar approach has been taken for parking assistance. In the first systems, the vehicle can detect the exact space available for parallel parking through ultrasonic sensors and compute the best trajectory. This trajectory is then executed through a total control of the steering wheel. However, the driver who is responsible for the maneuver must control the speed of the vehicle and avoid any possible collision. New systems under test use cameras to map the visible environment of the vehicle, find a parking spot if possible (in any configuration such as in a parking lot), and then execute the trajectory, with the driver in the loop for monitoring the maneuver and abort it if any danger occurs.

A particular case of assistive function that does not need the driver's intervention is now gaining attention: It is the collision avoidance or collision mitigation. This function is usually an extension of the regular ACC but it may also involve the use of vision sensors



■ Fig. 60.4
Lane-keeping assistance by camera (Valeo)

and complex fusion of data from the radar and the camera to be certain of a collision risk. In the mitigation function, if the sensors and the risk assessment of the system detect with certainty that a collision is going to happen, it applies the brakes without human intervention in order to reduce the severity of the crash. In the collision avoidance, it will start by a warning to the driver of incoming danger, then if the danger increases will start applying reduced braking to finish with full braking (and various restrains to the occupants) when the collision is certain to occur (Hillenbrand et al. 2006).

Autonomous driving. In this class, we find intelligent functions that allow the vehicle to drive completely by itself without any human intervention. Some authors prefer to speak of automatic driving since in some cases, the vehicle is in communication with its environment or with other vehicles and therefore is not “autonomous.” In any case, these functions must still be performed under the control or supervision of a human driver for public roads since this is required by the legislation in almost every country (Vienna convention of 1968). This need for a human supervision is also a requisite of the car manufacturers that are still reluctant to take full responsibility for accidents that may occur. If we accept this human supervision, then we can talk of “highly automated driving” and such intelligent functions are being developed and tested in the framework of the HAVE it European project (www.haveit-eu.org).

On the other hand, it is clear that the functions described previously can easily extend themselves to fully automatic driving in simple environments where the risk of accident

is minimal. This is particularly the case when the speed of the vehicle is very low. Such is the case for parking environments where a function of “valet parking” could be deployed in specific parking lots (private and hence not subject to the Vienna convention) but without any driver on board. The technology has already been demonstrated by researchers but seems difficult to deploy in traditional vehicles and parking due to the chicken-and-egg problem: Who starts first, the parking or the automated application? However, we will see later that the solution might stand with vehicle fleets such as car-sharing fleets.

Fully automatic vehicles have seen a large boost through the three DARPA Challenges (2004, 2005, 2007). These challenges put several hundreds of research teams in the field of automated vehicles, although only a few made it to the final competitions. In 2004, no team was able to complete the course but in 2005, three teams were able to complete a course of more than 200 km in the desert at close to 30 km/h average speed. In 2007, the Challenge was called the Urban Challenge because the course was in an urban-like environment and it included several vehicles (automated and nonautomated) present at the same time on the track. Although most of the vehicles were overloaded with sensors and computer equipment, at least one vehicle (from Ibeo) had nothing visible to distinguish it from a standard road vehicle, which means that the technology is quite close to the market. This is probably why Google has recently invested in experiments on fully automated vehicles driving in California (► Fig. 60.5).

Another striking experiment with fully automated cars (perhaps a little less advertised) was also conducted by the University of Parma between Italy and Shanghai in the summer



■ Fig. 60.5

Stanley from Stanford University, first winner of a DARPA challenge (DARPA)



■ Fig. 60.6
Vislab automated vehicle (University of Parma)

of 2010 (<http://viac.vislab.it/>). Two electric vehicles drove 13,000 km mostly in automated mode for the first one (a driver was supervising) and the second one in fully automated mode, but following the first one (using a communication link to pass localization information) (● Fig. 60.6).

So, are we at the beginning of a new era for automated vehicles or at least of vehicles with several intelligent functions to improve safety and make life simpler and more comfortable for drivers as well as passengers? The answer to this question is still far from simple and will mostly depend on the market forces and the acceptance of the public for these new functions. With a market of more than 60 millions of vehicles sold worldwide each year, the massive introduction of these new functions would mean very low cost (as we have already seen with the dissemination of antilock systems).

Research by the Dutch Ministry of Transport (RESPONSE 2 2004) has elaborated eight groups of factors that influence acceptance according to the definition above:

- Price. The cost for the customer can be reduced by incentives at the introduction stage.
- Benefit. It is important here to mention the customer's perception of the benefit. Issues such as risk compensation may influence this benefit perception. Therefore, communication is crucial.
- Introduction scenario. When the system is introduced to the market, the demonstrator must be credible and trustworthy. Transport authorities and user representation groups (such as ADAC, ANWB, AAA) and companies with strong positive image as to innovation may bring to bear strong benefits on user acceptance.

- Level of controllability. Controllability is an essential issue. The driver must remain in control, and must perceive it this way. To make sure that a system will not be “overruled” too often, there should be sufficient confidence by the driver in the proper functioning of the system. Bad experiences decrease this level of confidence.
- Image. A system may provide a certain status to the user. Corporate identity for fleet owners could play a role.
- Ease of use. Before introduction or first use, the user must be well informed about the system’s limits, human–machine interface, and performance. During driving, the system should give feedback on its performance (“situational awareness”).
- Comfort. The system should enhance the driver’s comfort, by workload reduction. The HMI design is an important enabler.
- “Joy of Controllability.” This issue refers to the pleasure that many drivers have when driving. It is greatly dependant on the function for which the vehicle is used (commercial, private, etc.)

3 Public Urban Vehicles

The last 10 years have seen the rapid development of car-sharing systems in Europe, Japan, and North America (Shahen 2010). Such systems are based on the assumption that the residents of large cities need a private automobile only very occasionally and that other modes such as walking, cycling, and mass transit are often more convenient, cheaper, and faster than the private car. Furthermore, the private car is often (and rightly) perceived as expensive and troublesome especially when parking is concerned. The facility (and/or cost) to find a parking space at the destination is often the key element in choosing this mode of transport.

However, up to now, the need to have, even occasionally, a car available 24 h a day is sufficient to justify the cost of purchasing, maintaining, and parking (often more than 95% of its time) a private automobile.

Car-sharing organizations (public or private) now bring a solution to this need through a registration to a service where you can pick up a “public vehicle” close to your location without any human intervention. Two types of service are now available: closed-loop systems and open-end systems. In the closed-loop system, you must return the car to its point of origin while in the open-end system, you can return the vehicle in any station. Some systems such as those offered by VuLog (www.vulog.fr) allow to leave the car anywhere in certain perimeters, the cars being located by GPS for pickup. Reservations can be made by telephone, smartphone, or Internet up to the last minute but sometimes are not needed.

The cities are often interested in the deployment of car-sharing system because this is a solution to the lack of parking space: It is considered that each car-shared vehicle can replace up to 20 private vehicles. Furthermore, these vehicles can be much cleaner than the private ones they replace and that are often very old ones and hence very polluting. They can even be electric as in the case of CitéVU in Antibes and Nice and Autolib in Paris.

Are these cars “intelligent”? Most of the time, not more than any standard vehicle but the system in itself can be considered as “intelligent” in the sense that it uses sophisticated management techniques and relies heavily on localization of the vehicles and digital communications.

However, with the deployment of these car-sharing systems in a larger scale, the operators are starting to look for more specific vehicles that are better adapted to the service than the traditional ones (that have to be fitted with specific electronics for the service). Those new vehicles are bound to introduce more intelligence for two principal reasons: to make the operation of the service more economical and to make the vehicles more attractive to the public and to the cities.

To make the vehicles more attractive to the public, we have to look at the previous factors that influence the decision of the user, except we are not bound anymore by the investment price (only the service price). So, we have to look for ease of use, controllability, comfort, and image. The image factor for such a service is probably key and may very well attract the younger population if the vehicles have this “high-tech” image brought by intelligent vehicles as long as it is easy to use!

To make the vehicles attractive to the city, the main issue is to make them acceptable to the population (even if they do not use them) and to give a good image “clean-tech” to the city. In this sense, electric vehicles such those offered in La Rochelle or Nice by Veolia are definitely a plus (► [Fig. 60.7](#)).

Finally, for the service operator, the key elements are the attractiveness of the service (and hence its cost but also the other factors presented above) and the cost of operation. To reduce the cost of operation, one key element is the safety of the vehicle to reduce the



■ Fig. 60.7
Autobleuve car sharing in Nice (Veolia)



■ Fig. 60.8

ICVS car-sharing system from Honda in Motegui (Honda)

maintenance and insurance cost. In this sense, intelligent vehicles can lower this cost if they include speed limitation, obstacle avoidance, and parking aids.

The operation of open-end car-sharing service could also be made more efficient if we could find a solution to move the vehicles automatically between stations. In 1994, INRIA has presented a technique to move several vehicles with a single driver using platooning techniques (Parent et al. 1995). This technique was developed for the Praxitele project that was the first demonstration of a large car-sharing system using ITS technologies such as localization and communications and electric vehicles. However, the platooning technique was not put in operation because of regulations on public roads although it was presented by Honda in 1998 in Motegui to demonstrate a car-sharing system with automated parking and platooning (● Fig. 60.8).

Nevertheless, this platooning technique, associated with automated parking and recharging at stations, is now considered for the Cristal vehicles that are being developed by Lohr Industries (responsible for the Translohr tram and the NeoVal automated metros with Siemens). These vehicles can definitely be considered as among the most “intelligent” vehicles under development (<http://www.cats-project.org/>) (● Fig. 60.9).

4 Urban Freight Delivery

Urban freight delivery is both a contributor to and a victim of the growing congestion in urban areas, as it exposes the population to noise, pollution, and nuisances. If no measures are undertaken in the future, statistics show the risk of a continuous increase in traffic



■ Fig. 60.9
Cristal vehicles in platoon (Lohr Industries). Design sketch (Lohr)

volumes that will be due in part to freight flows (about 20%). Such a situation affects the quality of life as well as the environment, and means a loss of efficiency for the freight transport itself.

Today's solutions are often based on restrictive policies that include low emission zones, access control, road pricing, or time limits for the logistic operations. It is only in the last few years that experimental initiatives have been going toward a positive approach, in which public authorities offer ad hoc facilities like freight villages or reserved lanes or last mile deliveries using electric and intelligent vehicles.

These last-mile electric vehicles, associated with a transshipment platform, have been put in place in several European cities, thanks to the Elcidis European Project (www.elcidis.org). In La Rochelle, the service is now continuing with the company Proxiway (subsidiary of Veolia). The electric vehicles are constantly tracked by GPS and optimized in their operation (► Fig. 60.10).

Between 2005 and 2008, the 13 partners of the FIDEUS project (http://ec.europa.eu/research/transport/projects/article_5013_en.html), coordinated by Centro Ricerche FIAT and cofunded by the European Commission, developed a new approach for the freight delivery in urban space by proposing a family of vehicles with high performance, a reorganized logistic flow and a telematic tool for the logistics management.

The benefits expected are social (less congestion/environmental effects due to freight delivery) and economics (better efficiency in the operations). In terms of policy, the public authorities will have a greater degree of freedom in traffic control, with minimal effects on the operators.

From a practical point of view, the FIDEUS project developed a family of vehicles with high performance in terms of environmental impact reduction, noise level control, and



■ Fig. 60.10

Elcidis electric vehicle in La Rochelle (Elcidis)

ergonomics. The basic idea is to exploit the different features of these vehicles to achieve an efficient logistic flow toward the cities. A specific strategy will be elaborated to move freight into city centers with fewer trips by medium-to-large vehicles and to deliver the parcels using a micro-carrier that is able to circulate in pedestrian areas without any restriction. An alternative is a van with an ad hoc adaptation that could carry out deliveries in urban zones where low emissions and noise levels are mandatory.

This approach requires some cross-solutions to enhance and complete the capabilities of the proposed set of vehicles. For this purpose, FIDEUS has identified a multimode container to facilitate the freight handling and delivery, and a telematic system to manage the logistic flow. Obviously this extended package will adopt other practical measures, for example, to achieve easy loading/unloading operations, to have transshipment areas or reserved lanes, to enable the vehicles to exchange data, to track the goods, etc. (► [Fig. 60.11](#)).

5 Rapid Transit (Including BRT, PRT, and CTS)

With the problems of congestion, parking difficulties, emissions, and high cost of private vehicle usage (not to mention access restrictions which are becoming more and more frequent), public transit is regaining popularity and usage in large cities throughout the world.



■ Fig. 60.11
Fideus microvehicle for pedestrian zones (Fideus Project)

In order to improve image as well as economic efficiency, transit vehicles are turning to ITS and become themselves intelligent vehicles. The first step that has been taken in the late 1990s and early 2000s was to introduce localization and communication technologies. It can be said now that most buses running in advanced cities are now equipped with such technologies and that the operator uses them to optimize their operation through real-time regulation systems.

In order to make buses safer, the operators are also turning to the same technologies as the automotive sector with the introduction of various sensors such as radar, lidar, ultrasounds, and vision. Lateral guidance is now available with several manufacturers using technologies based on vision (Irisbus/Siemens), magnets (Phileas and Miller 2009), or GPS (Cheng et al. 2007). Such techniques allow the buses to use less space with better safety and are strongly considered for the implementation of BRT (Bus Rapid Transit) lines. Another advantage of lateral guidance is the possibility to do precision docking, allowing boarding for wheelchairs and strollers (🔗 Fig. 60.12).

Fully automated buses have also been considered using lateral guidance plus longitudinal guidance. In this case, the driver can be fully eliminated or his role reduced to supervision. The main advantage of longitudinal control is to implement a better regulation of the bus schedule but also to improve the comfort, the safety, and the mileage. Such buses were demonstrated in Aichi by Toyota during the World Expo of 2005. Platooning techniques were also used to increase the throughput. In the USA, the Smartbus project also developed similar techniques (Smartbus 2000) (🔗 Fig. 60.13).

For a long time (since the middle of the twentieth century), it has been considered that the optimal public transit would be a fully automated system with small vehicles running on a dedicated network of tracks with direct origin–destination without intermediate stops. These systems that have seen several demonstrators built are called in the literature PRT for



■ Fig. 60.12

Precision docking with Irisbus in Rouen using vision from Matra-Siemens (Parent)



■ Fig. 60.13

IMTS Buses from Toyota in Aichi World Expo (Toyota)



■ Fig. 60.14
ULTra vehicles at Heathrow (ULTra PRT Ltd.)

Personal Rapid Transit. Three of these demonstrators are now operational although none in revenue service: one test track using mechanical guidance by Vectus in Sweden (hence we will not consider them as intelligent vehicles) and two using fully automatic road vehicles at Heathrow (ULTra) and Mazdar (2GetThere) (► [Figs. 60.14](#) and ► [60.15](#)).

However, these PRT suffer from the high cost of the infrastructure that has to be protected from intrusion, the high cost of the vehicles that are produced in very low volumes, and the low capacity compared to buses, trains, or metros.

The solution may come from CTS or Cybernetic Transportation Systems that are based on highly intelligent urban vehicles (called cybercars) with provision for fully automatic driving in nonprotected environments at low speed and capable of higher speeds in more protected ones. These vehicles have been developed by a number of



■ Fig. 60.15

2GetThere vehicle in Mazdar (2GetThere)

research institutes and companies since the early 2000s through several European projects such as CyberCars, CyberMove, CyberCars-2, CyberC3, CityMobil, and CityNetMobil (see www.cybercars.org).

In a CTS, a fleet of highly intelligent vehicles (potentially from different suppliers) form a transportation system for passengers and for goods linked to existing mass transit and operating as a complement to the other modes. It can be seen as an evolution of the car-sharing concept with highly intelligent vehicles. The fully automated mode could be reserved to specific areas of the network while in other areas some of the vehicles (called dual mode) could be operated in manual mode with a driver (public mode as with Cristal) or in self-service mode by the user.

Such vehicles have already been presented in various cities during the CityMobil and CityNetMobil projects for “showcases” and in particular during a 3-month test in the city of La Rochelle by INRIA (● Fig. 60.16).

A similar system has also been developed in 2002 by Toyota for a demonstrator in their showcase of Tokyo with fully automated urban vehicles but on a dedicated track, therefore closer to the PRT concept. However, it would not be difficult to consider that these vehicles could run automatically at low speed in nonprotected environments or in manual mode (● Fig. 60.17).

Perhaps this is the long-term future of the automobile: fully automated low-speed modes in the densest parts of the city, fully automated high speed modes on dedicated and



■ Fig. 60.16
Cybercar from INRIA (Parent)



■ Fig. 60.17
Automated e-Cars from Toyota in Tokyo showcase (Toyota)

protected infrastructures, and strong assistance elsewhere and in any case, with connectivity with the other transportation modes through onboard system and personal assistants. Most of the cars in cities would be in car-sharing mode to take advantage of the complementarities with the other modes.

6 Long-Distance Freight

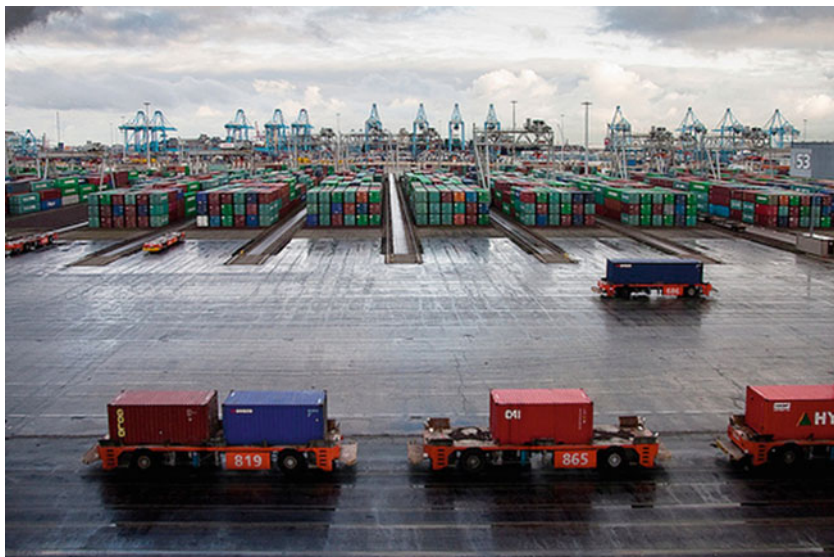
Trucks for long-distance freight transport have been the first vehicles to benefit from the technologies of intelligent vehicles because this is where the case for cost-benefit ratio is the clearest. Indeed, the truck fleet operators have soon discovered that the benefits in terms of safety can easily outweigh the costs since a disabled truck (if only for filling an insurance claim) can lead to large costs in delayed delivery or loss of load.

The first systems to be installed on board of truck were therefore linked to lane departure and collision warnings. They are based respectively on vision and radar (or lidar) sensors. The latest ones can also be used to implement ACC that brings a comfort feature to the driver (but the major benefit is in term of safety).

To progress toward more intelligence and more automation, several research projects on truck platooning have been conducted in Europe as well as in the USA. The main advantage seen in platooning concerns the reduction of the aerodynamic drag that can improve the mileage by as much as 20% according to several studies. The possibility to remove the driver from the control loop (whether he/she stay in the cabin) has been



■ Fig. 60.18
Trucks in platoon mode (Chauffeur Project)



■ Fig. 60.19
Automated vehicles in Rotterdam container terminal (ECT)

considered but seems difficult to implement in the medium term unless special roads are built (although fully automated trucks such as the container handlers in harbors already work in industrial environments) (► Figs. 60.18 and ► 60.19).

References

- Cheng P-M, Donath M, Gorjestani A, Menon A, Newstrom B, Shankwitz C (2007) Advanced BRT volume II: innovative technologies for dedicated roadways, Report no. CTS 08–07. University of Minnesota, Minnesota
- Hillenbrand J, Spieker AM, Kroschel K (2006) A multilevel collision mitigation approach: Its situation assessment, decision making, and performance tradeoffs. *IEEE Trans Intell Transp Syst* 7(4):528–540
- Miller MA (2009) Bus lanes/bus rapid transit systems on highways: review of the literature. California PATH Working Paper UCB-ITS-PWP-2009-1
- Parent M, Daviet P, Abdou S (1995) Platooning for small public urban vehicles. Fourth international symposium on experimental robotics, ISER'95 Stanford, 1995
- RESPONSE 2 (2004) Final report: ADAS – from market introduction scenarios towards a code of practice for development and evaluation. EC Contract Number IST 2001–37528
- Shaheen SA, Cohen AP, Chung MS (2010) North American carsharing. 10-year retrospective. *J Transp Res Board* 40(2010):35–44. Transportation Research Board of the National Academies, Washington, DC
- Williams B (2005) The CALM handbook. ISO TC204: ETSI ERM TG37 Publication

Index

A

AAP. *See* Active accelerator pedal

AASHTO Green Book, 1139, 1140

Abbreviated injury severity (AIS), 872

ABS, 1557

ABS/ESP System, 255

ACC. *See* Active cruise control; Adaptive cruise control; Autonomous cruise control

Accelerator force feedback pedal (AFFP), 254

Accelerometer, 35, 443–446, 449–451

Acceptance, 118–136

- studies, 650–654

Accidents, 118, 124, 128–130, 133, 134, 136

- research, 730, 731

- simulation, 809

- statistics, 730

ACC state management, 620–625

Accuracy, 439, 441–443, 445, 451–453, 455–461

Ackermann's model, 1367

A control theory model, 607

Acoustic sensor network, 425

Acoustic warning, 669

Active accelerator pedal (AAP), 599

Active cruise control (ACC), 1062

Active gas pedal, 670

Active pedestrian safety, 787–789, 791, 796–826

Active safety, 661, 666, 710, 711, 715–717, 722–726

Active sensors, 1036–1038, 1063, 1064

Actuating elements, 787, 796–798, 805–808, 810, 814, 819–821, 826

Actuations, 520, 524–525

Actuation systems

- ABS, ESC and ACC systems, 50–52

- brake-by-wire systems, 53–54

- energy source category

- electrical actuators (*see* Electrical actuators)

- mechanical actuators, 45

- piezoelectric actuators, 49

- pneumatic and hydraulic actuators, 48

- thermal bimorphs, 50

- highly automated vehicle

- autonomous vehicles, 55, 56

- under development system, 55–58

- sophisticated applications, 44

- steering and steer-by-wire systems, 52–53

Adaptability, 515–516, 520

Adaptive and cooperative cruise control

- controller design

- control problem formulation, 198–199

- control system structure, 199, 200

- fuzzy logic control, 205

- linear, 202–203

- MPC, 204–205

- nonlinear, 203, 204

- spacing policy, 200–201

- string stability, 201

- production-level systems, 192

- system architecture and operation modes

- ACC/CACC controller, 196

- HMI, 196–197

- normal vehicle operation, 194

- range sensor, 196

- transition state machine, 194, 195

- vehicle states sensing, 196

- wireless communication, 198

- wireless communication link, 194

- traffic flow characteristics, 193

Adaptive brake assist, 669, 671

Adaptive cruise control (ACC), 18, 19, 21, 24, 91–93, 100, 101, 108, 497, 498, 511, 512, 514, 515, 526, 532, 615–654, 662–665, 671, 674, 679–683, 685, 1062

Adaptive dampers, 669, 671

Adaptive headlights, 526–528

ADAS. *See* Advanced driver assistance systems

Ad-hoc vision algorithm, 1066

Advanced driver assistance systems (ADAS), 8,

- 118–128, 133–137, 495, 497, 511–517, 519, 520,

- 522–526, 529–532, 712, 715, 716, 723, 1002, 1522,

- 1525–1529, 1531, 1532, 1535, 1557, 1558

Advanced vehicle speed adaptation system (AVSAS), 590–594

Advertised services, 1243–1247

Advisory, 585, 587–590, 596–598, 600

AFFP. *See* Accelerator force feedback pedal

Age, 500, 503, 506, 507, 509, 522, 530

AHS. *See* Automated highway systems

- AIMSUN, 595
 - Airbag, 669, 674
 - Air-fuel ratio, 288
 - Alert application, 1123
 - American Automobile Association (AAA), 943
 - Amplitude modulation (AM), 961
 - Anonymity, 1182, 1183, 1186, 1191, 1196, 1198, 1203
 - Anonymous in-vehicle data, 723
 - Anticollision, 37
 - Antilock braking system (ABS), 50, 51
 - Application, 1219–1225, 1227, 1228, 1230–1236, 1238–1245, 1247–1265
 - Architectures for intelligent vehicles, 1278–1286
 - Area control, 109–112
 - Area hazard warning, 1240–1243
 - ARENA, 595, 599
 - Around View System, 844, 850, 854, 856
 - Artificial intelligence, 84, 88–89
 - Artificial neural network (ANN), 958
 - Assist torque, 693–695, 697, 701, 703, 706
 - Asymmetric cryptography, 1190
 - Attribute uncertainty, 73, 74
 - Audible warning, 701
 - Auditory, 585, 586, 597
 - Authentication, 1175–1215
 - Authorization, 1224–1225, 1233, 1234, 1236–1240, 1242, 1246–1251, 1253–1257, 1261, 1262
 - authority, 1237, 1256, 1262
 - statement, 1224, 1234, 1237, 1238
 - Autobleue, 1565
 - Auto21 CDS, 89, 93, 95, 96, 112, 113
 - Auto exposure, 1017
 - AutoExposure control, 1016
 - Autolib, 1564
 - Automated buses, 172, 184–188
 - Automated highway systems (AHS), 83, 96, 112, 113, 169, 188
 - Automated lane keeping, 691, 694, 697
 - Automated vehicles, 18, 25, 26, 28, 168–173, 185, 187, 189
 - Automatic braking, 211, 764, 773, 774
 - Automatic control, 506, 511, 512, 514–515, 531, 585
 - Automatic crash notification (ACN) system
 - crash characteristics, 866
 - eACN systems
 - data-packet, 868
 - GSM SMS technology, 868
 - implementation, 869
 - in-vehicle sensor systems, 866
 - rescue sequence, 866, 867
 - trauma triage decision making, 877
 - moderate-to-high severity crash, 866
 - principal components, 867, 868
 - Automatic emergency braking, 666, 670, 671, 673, 677, 682
 - Automatic evasion assistance, 761, 763, 764, 771, 773–780
 - Automatic exposure control (AEC), 1006
 - Automatic partial braking, 670
 - Automation, 606, 609, 610, 1522, 1526–1534
 - Autonomous/adaptive cruise control (ACC) system, 51
 - Autonomous braking system, 805–806
 - Autonomous control, 506, 513, 515
 - Autonomous cruise control (ACC), 1062
 - Autonomous driving, 238, 252, 258–260, 265, 1522–1536, 1561
 - Autonomous parking, 831, 846, 859
 - Autonomous vehicles, 1023
 - Availability, 439, 443, 461, 1223–1224, 1231, 1233, 1234, 1236–1244, 1246, 1247, 1249–1252, 1257
 - AVECLOS, 899
 - Averaging, 375, 376
 - AVSAS. *See* Advanced vehicle speed adaptation system
- ## B
- Basic design of safety critical systems, 277–278
 - Battery electric vehicle (BEV), 299, 300
 - Bayesian formulation
 - Markov process, 1387
 - measurement and dynamic model, 1388
 - posterior distribution/Bayesian belief, 1387
 - probabilistic model, 1388
 - recursion equation, 1388
 - Bayesian occupancy filter (BOF), 1410–1412
 - Behavior modeling and learning
 - Bayes filter
 - advantage, 1314
 - belief state, 1315
 - Markov hypothesis, 1314
 - probabilistic estimation, dynamic system, 1313
 - transition probability, 1315
 - variables, 1314
 - experimental data, 1327–1328
 - GHMM
 - analysis, 1321–1332
 - continuous state spaces, 1320

- definition, 1320
- model's structure, 1324
- parameter learning, 1324–1325
- probabilistic model, 1321–1322
- sampling_step, 1328–1329
- state space discretization, 1321
- topological map update, 1322–1324
- HMM, 1318–1320
- leeds data parameters, 1328, 1329
- model learning
 - complex factors, 1316
 - intentional models, 1317
 - Kalman filter, 1315
 - metric models, 1318
 - modeled state, 1317
 - static/dynamic components, 1317
 - tools, 1317
 - train station and trajectories, 1316
 - transition probability, 1315
- model size, 1329–1331
- motion pattern learning system, 1325
- motion planning algorithms, 1312
- motion prediction and state estimation, 1312–1313
- multisensor tracking techniques, 1326
- performance metrics, 1326–1327
- prediction accuracy, 1329, 1330
- prediction suitability, 1326
- reliable sensor fusion, 1326
- structure learning, 1326
- unsupervised learning, 1326
- Behavior-related risk, 1502
- Behavioural adaptation, 118–136
- Benefit calculation, 808, 809, 820–826
- Biases, 723
- Bicycle model. *See* Ackermann's model
- Bicycle motion model, 1396
- Binary Bayes filter, 387
- Binning, 1016
- Bin-occupancy density, 390
- Blind spot
 - assistant, 742, 749–750, 754
 - detection, 504, 511, 526, 528, 532, 1000
 - detector, 751, 754
 - information system, 742, 745–747, 751–752, 754
 - monitor, 747, 748, 754
- Blind spot information system (BLIS), 742, 745–747, 751–752, 754
- Bluetooth, 1094, 1114, 1120
- Bootstrap Trust, 1253, 1254

- Bot dots, 698
- Botts' dots, 698, 699, 704
- Brake assist system, 511, 526, 527, 787, 805
- Brake force feedback pedal, 254–255
- Brake functions, 243, 244, 246, 249
- Brake pedal, 239, 240, 242, 244–252, 254–258, 278, 280
 - characteristics, 246–247, 249, 250
- Braking distance, 762–764
- Braking intervention, 760, 769, 772
- Braking jerk, 668, 670, 673, 681
- Brightness, 1017
- Broadcast, 1096, 1099, 1103–1105, 1109, 1110, 1115–1117
- Brownian motion model, 1396
- Bus rapid transit (BRT), 184, 1556, 1568–1574
- By-wire approach, 52

C

- CACC. *See* Cooperative adaptive cruise control
- Camera, 366, 368–373, 375, 377, 378, 380, 388, 389, 392, 394, 464, 468, 480–482, 751, 752, 756
- Camera technologies
 - capture parameters, 1013
 - classification, 1013
 - definition, 1012
 - exposure parameters, 1016–1017
 - image format, 1016
 - mechanical issues, 1019
 - optics, 1017
 - sensor
 - CCD, 1017–1018
 - CMOS, 1018
 - color information, 1014
 - dynamic range, 1014–1015
 - geometry, 1013–1014
 - lenses, 1014
 - photolithographic techniques, 1013
 - software, 1019
 - speed, 1015–1016
 - time behavior, 1013
 - vision system, 1012
- Camera/video imaging systems (C/VISs), 561, 574, 578
- Carbon balance method, 288
- CAR 2 CAR Communication Consortium (C2C-CC), 44
- Car navigation, 464–486
- Car pooling, 1547–1549
- Carrier information signal, 319, 321

- Carrier phase tracking
 - code synchronization, 333
 - parameter value, 331
 - phase information, 332
 - signal component, 331, 332
 - transformation, loop, 331
- Carrier sense multiple access (CSMA), 1103–1105
- Car sharing, 1547–1549, 1564–1566, 1574
- Case-based reasoning, 88
- CCH. *See* Control channel
- Cellular automaton (CA) model, 547
- Centralized filtering, 456, 457
- Certificate authority (CA), 1189, 1192, 1197–1206, 1208, 1209, 1214–1215
- Certificate management, 1175, 1186, 1190–1207, 1214, 1215
- Certificate revocation, 1190, 1192–1197
- Charge-coupled device (CCD), 1017–1018
 - sensor, 842, 843
 - vision system, 39
- Chauffeur project, 1574
- CitéVU, 1562
- Clothoid, 369, 374–376
 - function, 762
 - model, 376
- Clustering algorithm, 590
- CMS-braking, 670
- Cognitive behavior, 499
- Cognitive perception, 505
- Collision
 - avoidance system, 685, 715, 717, 721, 726, 760, 762, 776–778
 - free path, 369
 - mitigation, 715, 765, 776, 780
 - risk assessment, 773
 - risk estimation
 - behavior recognition and modeling, 1487–1489
 - chosen implementation, 1492–1494
 - conformal transformation, 1495, 1496
 - Gaussian process deformation model, 1490–1492
 - implementation issue, 1492
 - one vehicle trajectory, 1499, 1500
 - overall architecture, 1485, 1486
 - probability distribution, 1496, 1497
 - property, 1492
 - real-world coordinates, 1497, 1498
 - risk aggregation, 1500–1502
 - risk with driving behavior, 1502, 1503
 - velocity, 821–825
- Color depth, 1016
- Color segmentation
 - color camera, 1044
 - HSV/HLS color space, 1046
 - illumination conditions, 1045, 1046
 - pixels, 1047
 - RGB and YUV space, 1046
 - thresholding, 1046
- Combinatorial certificate, 1192–1197, 1207
- Combined function, 243–246
- Comfort deceleration, 1063
- Commercial long-haul trucks, 1153
- Commercial motor vehicles (CMV), 990
- Common mode errors, 440, 441
- Communication, 466–468, 470, 472, 474, 477–478, 485, 486
- Complementary metal oxide semiconductor (CMOS), 1015, 1018
 - sensor, 842, 843
 - vision system, 39
- Components for vehicle navigation functions, 1278–1286
- Compression ignition (CI) engine, 295, 301
- Conditional random fields (CRFs), 1418–1419
- Confidentiality, 1175, 1181, 1182
- Confidentiality, authenticity, integrity, non-repudiation, 1221–1222
- Confounding, 723
- Congestion, 1540, 1542–1547
- Constant velocity model, 1396
- Continuity of service, 439, 443, 460, 461
- Continuous air interface long and medium range (CALM), 44
- Control, 83–113, 493–498, 502–509, 511–517, 519, 521–532
 - application, 1123–1125, 1129, 1135
 - design methods, 84–85, 112
 - functions, 513
 - loop, 717
 - of platoons, 83, 84, 89, 90
 - segment, 320–321
 - strategy, 249, 251, 262, 273–276
- Control channel (CCH), 1111, 1116, 1117, 1119
- Controllability, 696, 697, 771, 772, 780
- Cooperative adaptive cruise control (CACC), 100, 101, 108, 109, 150–151

Cooperative awareness, 1226, 1232–1241, 1245
 Cooperative intersection collision avoidance, 1152
 Cooperative operation, 694
 Cooperative vehicle-infrastructure systems (CVIS), 94
 Cornering lights, 813
 Corner module, 273–274
 Corrective evasion assistance, 761, 763, 764
 Countermeasures, 609
 Crashes, 583, 586, 587
 Criticality assessment, 764, 765, 776
 Crossing pedestrian, 765, 767, 774
 Cryptographic, 1221–1223, 1234, 1236, 1241, 1243, 1247, 1250, 1251, 1253, 1254, 1263
 CSMA. *See* Carrier sense multiple access
 CTS, 1556, 1568–1574
 CU-criterion, 665–667, 671, 679
 Curvature, 374–377, 380
 – model, 375
 Curvature Warning, 211
 Curve estimation, 380
 Curve speed warning (CSW), 888–889
 Customer acceptance, 733, 761, 770–771
 CVIS. *See* Cooperative vehicle-infrastructure systems
 Cybercars, 1452, 1571, 1572

D

Danger and risk analysis, 278–279
 DARPA, 1562
 DARPA Urban Challenge, 1452–1453
 Data acquisition system (DAS), 563–565, 578
 Data association, 370, 383
 Data exchange application, 1123, 1127–1130
 Data fusion, 896–910
 Data privacy, 1522, 1531, 1533–1536
 Data retrieval, 1391
 Daytime running lights, 813
 DCAM, 1019
 Dead-throttle, 586
 Deceleration, 589–591
 Decentralized environmental notification (DEN), 1240–1243
 Decentralized filtering, 456, 457
 Decimation, 1016
 Decision sight distance, 1125, 1129, 1130, 1134, 1139, 1140
 Decoupled brake pedal, 247–250
 Dedicated short range communication (DSRC), 42, 44, 1108–1120, 1130, 1142, 1153, 1177, 1180, 1182, 1209–1212, 1214

Degradation, 277–279
 Delaunay triangulation, 1323, 1324
 Demand responsive transit (DRT), 1551–1552
 Democratization, 741
 Denial-of-service attacks, 1224
 Designing driver assistance systems, 571–577
 Design process, 676–685
 Design rules, 609
 Detection and tracking of moving objects (DATMO)
 – cluster-based models, 1396, 1397
 – geometric models, 1398, 1399
 – grid-based models, 1399–1400
 – inference
 – Bayesian filtering paradigm, 1409–1410
 – BOF, 1410–1412
 – curse of dimensionality, 1401
 – data association, 1404–1406
 – data segmentation, 1404
 – dynamics update, 1403–1404
 – MCMC approach, 1408, 1409
 – model-and grid-based approach, 1401, 1402
 – non-parametric filters, 1401
 – object level tracking (*see* Object level tracking)
 – parametric filters, 1400–1401
 – Rao-Blackwellization, 1406, 1408
 – scaling series algorithm, 1408–1410
 – track constraints, 1402
 – track creation and deletion, 1403
 – track existence, 1402, 1403
 – measurement models
 – dynamics model, 1396
 – physical model, 1393
 – pseudo-position, 1393
 – virtual sensors, 1394–1396
 – need, 1385
 – pattern recognition
 – detection, classifier ensemble, 1420–1422
 – object and people detection, 1418–1420
 – vision-based approach, 1414, 1416–1417
 – problem statement
 – Bayesian formulation, 1387–1389
 – coordinate frames, 1387
 – sensor fusion
 – data integration, 1422
 – feature-based framework, 1423
 – feature extraction, 1424, 1425

- geometry, 1426
- ground MRF, 1428
- perception systems, 1422
- semantic fusion, 1424–1425
- setup vehicle, 1426
- threefold advantages, 1426
- sensors
 - laser range finders, 1390–1391
 - optical cameras, 1389–1390
 - radar, 1391–1392
- taxonomy, 1386
- Detection range, 627–630, 648, 649, 734, 744, 747–749
- Development process, 807–809, 820
- Digital signature, 1182, 1190, 1193
- Dilution of precision (DOP), 328, 329
- Discomfort, 606, 607
- Discretionary lane changing (DLC), 550–551
- Discriminant circle, 966
- Distance between polylines, 1026
- Distance control, 615, 627, 629, 638, 640
- Distance measurement, 831, 833–835, 857
- Distances of zero-crossings (DZC), 965
- Distraction, 589
- Distributed (networked) service, 1251–1252
- Dolphin, 89, 93, 95, 96, 112
- Doppler shift, 388
- Drive functions, 243–244
- Driver
 - alertness, 118
 - assistance, 493–532
 - types and levels, 510–515
 - assistance systems, 1522–1536
 - centric vehicles, 1290
 - distraction, 506–510, 583, 588, 589
 - impairment, 896, 900
 - input, 514, 515, 520, 524
 - overloading, 589, 600
 - perception-response, 495, 498–503
 - underloading, 589
 - vehicle-Environment control loop, 607–608
 - warnings, 806–807
- Driver-initiated evasion assistance, 761, 763, 771, 780
- Driver in the loop, 720
- Driver modeling
 - driver behavior models
 - acceleration operational models, 544
 - car-following model factors, 540–541
 - classification, 539, 540
 - cognitive decision-making processes, 544
 - congestion dynamics, 539
 - crash-free environment, 539
 - individual differences attributes, 541
 - information-processing/execution system, 544, 545
 - intelli-drive systems, 540, 544
 - lane changing tactical models, 544
 - literature assumptions, 543
 - situational environmental factors, 541–542
 - situational factors, 542–543
 - tactical route execution decisions, 540
 - execution-time horizons, 539
 - macroscopic approach, 538
 - mesoscopic approach, 538
 - microscopic approach, 538–539
 - operational stage acceleration models
 - CA model, 547
 - fundamental diagram, 548–549
 - generalized force model, 545
 - GHP model, 545–546
 - Gipps model, 546–547
 - IDM/IDMM model, 547–548
 - OVM, 545
 - S-K model, 547
 - stimulus, 545
 - trajectories, 549
 - Wiedemann model, 548
 - tactical stage lane changing models (*see* Tactical stage lane changing models)
- Driving, 493–497, 530–532
 - behavior/behaviour, 120, 494–496, 510, 525, 531, 532
 - realization, 1486
 - recognition, 1485–1486
 - control task, 608
 - corridor, 616, 629, 630, 633–636, 640, 642
 - simulators, 146
 - tasks, 495–498, 504–506, 511, 524
- Drowsiness, 896–906, 908–910
- Drowsy and fatigue driver warning system
 - alarm modality
 - auditory, 980–981
 - haptic/tactile warning, 982
 - visual display, 979–980
 - alarm timing, 983
 - alerting system, 978
 - commercially available systems, 993
 - countermeasures
 - driver education, 991
 - legislation/enforcement, 990–991
 - non-technological, 978

- rumble strips, 991–992
 - strategies, 992
 - design, 988–989
 - detection system, 977
 - physical and physiological conditions, 978
 - system reliability and sensitivity
 - decision thresholds, 984–986
 - false alarms, 984
 - graded (staged) alarm, 987–988
 - likelihood, falling asleep, 986–987
 - user acceptance and trust, 990
 - Drowsy and fatigue driving problem
 - characteristics
 - conditions and causes, 946
 - construed consequences, 946
 - drowsiness definition, 945
 - fatigue definition, 946
 - homeostasis, 945
 - NIH and NHTSA, 944
 - related accidents, 946–948
 - sleep/wake cycle function, 945–946
 - design and implementation challenges
 - automatic *vs.* manual activation and deactivation, 969
 - distraction, 969
 - evaluation and user acceptance, 969, 970
 - privacy provision, 969
 - robustness, 968
 - sampling rate, 969
 - stakeholders buy-in, 969
 - detection methods
 - advantages and disadvantages, 967–968
 - ANN method, 958–959
 - classification, 966–967
 - control actions, 948
 - driver physiological symptoms, 948
 - feature extraction (*see* Feature extraction)
 - multiple linear regression, 956
 - neural network, 959
 - pattern classification schema, 955, 956
 - post-processing, 964–966
 - predefined mathematical model, 955
 - preprocessing, 961
 - road curvature removal, 958
 - road horizontal geometry, 958
 - signal windowing, 960
 - steering adjustment intervals, 957
 - steering angle amplitude, 959
 - steering macro-corrections, 958
 - warning/alert/automatic vehicle control, 948
 - driver's physical and physiological condition, 948–949
 - FARS, 944–945
 - fatal accidents, 943
 - fatal crashes and fatalities, 944
 - NASS/GES data, 944, 945
 - North American consumer, 943
 - US Government and businesses, 943
 - vehicle state variables (*see* Vehicle state variables)
 - DRT. *See* Demand responsive transit
 - DSRC. *See* Dedicated short range communication
 - DVD, 467, 468, 471, 474–476, 480
 - 6D-Vision, 760, 774–776
 - Dynamic, 1462–1464, 1466, 1470, 1471, 1473, 1477
 - carpooling, 1549
 - curvature model, 375
 - event, 1126–1127
- ## E
- Earth-centered earth-fixed (ECEF) coordinate system, 447
 - Earth-centered inertial (ECI) coordinate system, 447
 - Eavesdropping, 1181–1183, 1201
 - eCall, 713
 - ECDSA. *See* Elliptic curve digital signature algorithm
 - Eco-driving systems, 885–886
 - Eco-routing. *See* Fuel-efficient routing
 - ECU. *See* Electronic control unit
 - Edge distribution function (EDF), 378, 379
 - EEG. *See* Electroencephalography
 - Effectiveness, 796, 805, 807, 820–824
 - Effectiveness and acceptance, 606, 609
 - EHB. *See* Electrohydraulic brake
 - EHCB. *See* Electro hydraulic combi brake
 - Elcidis, 1567, 1568
 - Electrical actuators
 - DC motors, 45, 46
 - electromagnets, 46, 48
 - rotational motion, 45
 - smart coolant pump, 46
 - solenoid valve, 46, 47
 - stepper motor, 46, 47
 - Electroencephalogram, 949
 - Electroencephalography (EEG), 898, 901, 905
 - Electrohydraulic brake (EHB), 256–258
 - Electro hydraulic combi brake (EHCB), 255–256
 - Electromagnetic clutch, 57
 - Electromechanical brake actuator, 57, 58
 - Electromechanical stabilization techniques, 1005

- Electronic accelerator pedal, 253
 - Electronic control unit (ECU), 50, 1001, 1002
 - Electronic horizon (eHorizon) system, 890
 - Electronic stability control (ESC) system, 51
 - Electronic stability programs (ESP), 1000, 1557
 - Electronic throttle, 253, 280, 281
 - Electronic toll collection systems, 1156
 - Electrooptical/infrared sensor (EO/IR), 428
 - Elliptic curve digital signature algorithm (ECDSA), 1234
 - EMB. *See* Full electro mechanical brake system
 - Emergency braking system, 760, 765, 772, 780
 - Emergency breaking, 814
 - Emergency steering assist, 772
 - Emergency stop, 805, 814
 - Emission, 588
 - Emitted power density (EPD) filter, 426
 - Empirical driving research, 562–563, 570, 577
 - Empirical mode decomposition (EMD), 961
 - Encounter time, 1098, 1120
 - Encryption, 1181, 1190
 - Energy and powertrain systems
 - battery electric vehicle, 299, 300
 - climate change and CO₂ awareness, 284
 - energy consumption
 - primary energy consumption, 303, 304
 - tank-to-wheel energy consumption, 301–302
 - energy storage systems, 293
 - fuel cell electric vehicle, 298, 299
 - hybrid electric vehicle, 296, 297
 - internal combustion engine vehicle, 295, 296
 - new propulsion systems, 294–295
 - research methodology
 - calculation model, 288–290
 - GPS tracking, 285
 - PEMS, 285
 - real-world in-car measurements (*see* Real-world in-car measurements)
 - representative real-world driving routes, 289, 291
 - tank-to-wheel simulations and evaluations, 285
 - vehicle power requirement, 291–293
 - zero emission propulsion systems, 284
 - Enforcement, 583, 584, 586
 - Engineering effect, 119
 - Enhanced automatic crash notification (eACN) systems
 - data-packet, 868
 - GSM SMS technology, 868
 - implementation, 869
 - in-vehicle sensor systems, 866
 - rescue sequence, 866, 867
 - trauma triage decision making, 877
 - Enrolment, 1254, 1255, 1261–1263
 - Environment, 1462–1467, 1469–1473, 1477
 - models, 369
 - perception, 760, 761, 765–768, 780
 - for safe drive by wire, 280
 - Epidemiological driving research, 562, 570
 - Ergonomics, 1001
 - Errors-in-variables (EIV), 381–383, 385
 - eSafety, 1231
 - ESP. *See* Electronic stability programs
 - EUDC. *See* Extra urban driving cycle
 - Euro NCAP Offset Crash, 722
 - European Automobile Manufacturers' Association, 610
 - European New Car Assessment Program (Euro-NCAP), 713
 - European Telecommunications Standards Institute (ETSI), 44
 - Evaluating driver assistance systems, 560–577
 - Evasion assist, 664
 - Evasion maneuver, 761, 767, 770, 772, 780
 - Evasion trajectory, 766, 771, 772
 - Evasive steering, 760, 762, 765, 770–773
 - Evita, 1231, 1264
 - Evolution of vehicle architectures, 1288, 1289
 - Excess speed, 590–592
 - Exhaust gas mass emissions, 288
 - Exhaust gas mass flow, 288
 - Expectation-maximization algorithm, 1320, 1326
 - Expiration, 1256, 1257
 - Explicit, 1224, 1232, 1237, 1252, 1256
 - Extended Kalman filter (EKF), 329, 377, 459, 1337
 - Extended targets, 380, 381, 390
 - Extra urban driving cycle (EUDC), 289, 291
 - Eye blink, 898, 899
 - Eye gaze estimation, 916, 920–921
 - EYEMEAN, 899
- ## F
- Facial behavior analysis, 919
 - Facial expression recognition, 916, 918–920, 937
 - Facial feature point tracking, 918–926, 932–934
 - False alarms, 978, 984
 - False alert, 1233
 - False positive warnings, 723
 - Fast clustering and tracking algorithm (FCTA), 1413
 - Fatality, 587

- Fatality analysis reporting system (FARS), 943–945
 - Fatigue, 896–910
 - Fault detection, 443, 456, 457, 460
 - Fault exclusion, 443
 - Fault isolation, 443, 456, 457
 - FCX-system, 663, 665, 666, 669
 - FDW. *See* Following distance warning
 - Feature-based map, 368–373, 394
 - Feature extraction, 1077–1079
 - degradation phases, 961–962
 - EMD steering signal, 961, 962
 - IMF₁, 961–963
 - SDIE, 962–964, 966
 - SDZC, 963–966
 - Feature tracking, 1079–1083
 - Feedback, 84–87, 91, 99, 100, 105, 106, 112, 496–499, 503, 505, 506, 521, 523, 528, 531
 - control, 84–86, 91, 112
 - to driver, 496, 515, 520–522, 528
 - loop, 724
 - Feedforward, 101, 107
 - control, 101
 - Fidelity of the simulation, 725
 - Field effectiveness, 723
 - Field of view (FOV), 1066, 1389
 - Field testing driver assistance systems, 574
 - FireWire, 1019
 - Focus of expansion (FOE), 1076, 1084
 - Following distance warning (FDW), 597
 - Force-multiplier denial of service attacks, 1224
 - Ford Sync, 1113
 - Foreseeable misuse, 697
 - Forward collision, 659–686
 - avoidance, 659–686
 - conditioning, 661, 685
 - mitigation, 661, 685
 - warning, 659–686
 - Forward looking safety systems, 716
 - Freedom Car program, 294
 - Free space analysis, 780
 - Frequency modulation (FM), 961
 - FSRA. *See* Full speed range adaptive cruise control
 - Fuel
 - cell, 293
 - cell powertrain, 298
 - consumption, 288, 587, 588
 - cost, 587
 - Fuel-efficient routing, 883–884
 - Full electro mechanical brake system (EMB), 258, 259
 - Full speed range adaptive cruise control (FSRA), 615, 618–622
 - Fully automatic parking, 846, 859
 - Fully autonomous navigation, 360
 - “Function,” 606, 610
 - Functional algorithm, 797, 802–805, 810, 821
 - Functional safety and availability, 272, 276–277
 - Functional specifications, 902, 904, 908, 909
 - Functional targets, 243–246, 260
 - Fuzzy logic, 88
- ## G
- Gain, 1017
 - Gamma, 1017
 - Gaussian mixture, 391
 - Gaussian process deformation model
 - canonical GP, 1490
 - conformal mapping, 1490–1492
 - Gazis, Herman, and Potts’ (GHP) model, 545–546
 - GDC, 468–471, 473, 481
 - Generalized force model, 545
 - Generic Hough transform (GHT), 1416–1417
 - Geocasting, 1240, 1241
 - Geographic coordinate system, 447–449
 - Geometrical information, 455
 - GHMM. *See* Growing hidden Markov models
 - GHP model. *See* Gazis, Herman, and Potts’ model
 - 24 GHz radar sensors, 743, 745, 747, 753, 754
 - GigE, 1019
 - Gipps model
 - operational stage acceleration models, 546–547
 - tactical stage lane changing models, 550
 - Global navigation satellite systems (GNSSs), 437, 439–444, 451–453, 455, 459–461
 - dead-reckoning, 437
 - deployed and under construction data, 315
 - differential
 - atmospheric and satellite errors, 337–338
 - augmentation systems, 338–340
 - signal measurement, 338
 - GLONASS, 315, 316
 - GPS satellites, 315
 - L1-, L2-and E5-signals, 316
 - position estimation
 - least-square (LS), 327–329
 - MMSE, 329
 - principles
 - positioning geometry (*see* Satellite positioning geometry)

- signal property, 318–320
 - system components and structure, 320–321
 - pseudorange and position relation, 322–323
 - pseudorange error sources, 333–335
 - pseudorange measurement
 - carrier and code phase tracking, 331–332
 - carrier phase tracking, 332–333
 - navigation data, 333
 - signal acquisition, 330–331
 - received signal and pseudorange relation
 - random distortion, 324, 325
 - systematic distortion, 325–327
 - undistorted signal, 324
 - receivers
 - antenna, 335–336
 - functionality, generic division, 335, 336
 - hardware pseudorange tracking block, 337
 - hardware signal acquisition implementation, 336–337
 - RF front end and ADC, 336
 - software, 337
 - technology limitations, 314–315
 - vehicle positioning, 314
- Global nearest neighbor (GNN), 422, 1405
- Global positioning system (GPS), 345, 347, 348, 353–360, 1112
- GMTI. *See* Ground moving target indication
- GNSS. *See* Global navigation satellite systems
- GPS. *See* Global positioning system
- Gradient operator sobel filter, 1077
- Green, 1237
- Grid mapping, 1400
- Ground-based augmentation systems (GBAS), 340
- Ground moving target indication (GMTI), 422, 423
- Growing hidden Markov models (GHMM)
 - analysis, 1321–1332
 - continuous state spaces, 1320
 - definition, 1320
 - model's structure, 1324
 - parameter learning, 1324–1325
 - probabilistic model, 1321–1322
 - sampling_step, 1328–1329
 - state space discretization, 1321
 - topological map update, 1322–1324
- Gyroscope, 444–446, 449–451
- ## H
- Hands-off detection, 696, 702, 703
- Haptic, 585, 586, 591, 595, 1557
 - communication, 581
 - mechanisms, 585
 - warning, 591, 700, 701
- Harris corners, 1079, 1084, 1086, 1087
- Have-IT Project, 1561
- Hazard warning lights, 670
- HDD, 464, 466–469, 471, 474, 476, 479, 486
- HDR sensors, 1015
- Heartbeat, 1186, 1188, 1209
- Heat-treated windshields filter, 1001
- Herzberg, 606
- Hidden Markov model (HMM), 1079, 1318–1320, 1487, 1488
- Hidden terminal, 1100, 1103–1106, 1120
- Hierarchical, 83–113
 - behavior, 608
 - control, 89–96
- High automation, 1529, 1534
- High-level controller, 199, 200
- High occupancy toll (HOT) lanes, 1128
- High occupancy vehicles (HOV), 1549
- Histogram methods, 1364
- Histogram of oriented gradients (HOG), 1417
- HMI. *See* Human-machine interface
- HMM. *See* Hidden Markov model
- Holistic approach, 713, 715, 718
- Hough transform, 379, 1050
- Hours-of-service (HOS) regulations, 990
- HOV. *See* High occupancy vehicles
- Hue, 1017
- Human aware navigation, 1472–1477
- Human factors engineering, 571–573
- Human machine
 - interaction, 609–610
 - interface, 495, 496, 520, 609, 620–625
- Human-machine interface (HMI), 196–197, 732, 742, 744, 760, 761, 770, 771
- Human reaction time, 499–503
- Human-vehicle
 - environment, 493
 - interface, 496, 520–525
- Hydraulic brake assist, 660, 669, 671
- Hypovigilance, 584, 588, 589, 600
- ## I
- Identifying user requirements, 571
- IDM. *See* Intelligent driver model
- IDMM. *See* Intelligent driver model with memory
- IEEE, 1103, 1111
- IEEE 1609, 1177, 1191, 1207
- IEEE Standard 1609, 1228

- Image geometry, 1016
- Image processing, vehicular applications
 - functionality, 1000–1001
 - industrial applications, 1000
 - lighting control, 1002, 1003
 - machine vision, automotive field
 - ADAS systems, 1002
 - camera features, 1003, 1004
 - environmental conditions, 1003
 - illumination conditions, 1006–1008
 - oscillation and vibrations, 1004–1006
 - vehicle ego-motion, 1004, 1005
 - technical feasibility, device positioning, 1001
 - video surveillance systems, 1000
 - wiring and positioning, 1002
- Image stabilization, 1005
- Implicit, 1224
- Implicit shape model (ISM) algorithm, 1416
- IMTS, 1570
- IMU. *See* Inertial measurement unit
- Inappropriate alarms, 978
- In-car navigation system, 437–439, 442, 453, 461
- Inertial measurement unit (IMU), 417–421, 445, 450–453
- Inertial navigation system (INS), 35
- Inertial sensors, 716
- Inevitable collision states (ICS)
 - collision avoidance, 1437
 - compactor scenario, 1438
 - C-space, 1437
 - definition, 1438
 - motion model, 1438, 1439
 - state-time space, 1438
- Information, 496, 497, 500, 501, 512, 513, 517, 519, 524, 530, 531
 - fusion, 437, 438, 446, 455–460
 - strategy, 733
- Informational application, 1123
- Informational/warning ISA, 593, 594
- Informative systems, 1557
- Infrastructural pedestrian safety, 792–793
- Injury risk curve, 822–825
- Injury risk function, 718
- Input module characteristics, 240–242, 271
- Instantaneous topological map (ITM) algorithm, 1321
 - edge adaptation, 1324
 - matching, 1323
 - node adaptation, 1324
 - properties, 1322–1323
 - weight adaptation, 1323
- In stochastic simulation, 725
- Insurance Institute for Highway Safety (IIHS), 713
- Integral safety, 710–727
- Integrated longitudinal control, 251–252
- Integrated safety, 519–520
- Integrated safety system, 664
- Integrated vehicle dynamic, 273–276
- Integrated vehicle safety, 796–798
- Integrity, 439, 442, 443, 451, 455, 460, 461
- Integrity and reliability of map-matching, 356–360
- Integrity monitoring, 442–443
- IntelliDrive, 1175, 1177, 1181, 1190, 1192, 1207–1208
- Intelligent assistance systems, 606
- Intelligent driver model (IDM), 148, 149, 547–548
- Intelligent driver model with memory (IDMM), 547–548
- Intelligent functions, 1556, 1557, 1561, 1563
- Intelligent speed adaptation system (ISA), 498, 519
- Intelligent transportation systems (ITS), 6, 33, 34
- Intelligent vehicle highway systems (IVHS), 83, 91, 92, 95–96, 110, 112, 113
- Intelligent vehicles, 83, 112
 - basic definitions
 - ADAS (*see* Advanced driver assistance systems)
 - autonomous, 7–8
 - cognitive and motor actions, 8
 - eco driving mode, 8
 - intelligent definition, 7
 - vehicle dynamics, 9
 - comfort and safety driving, 9
 - communications systems, 11
 - driver assistance, 11
 - energy and environment, 4–5
 - fully autonomous driving, 11–12
 - global positioning perspective, 11
 - governmental efforts
 - automation and advanced technologies, 5
 - environmental pollution reduction, 5
 - intellidrive program, 7
 - ITS architecture, 6–7
 - ITS program, 6
 - longitudinal and lateral control, 10
 - OEMs, 12
 - road vehicles, 9–10
 - safety and comfort systems, 11
 - significant progress, 9
 - special vehicular systems, 10–11
 - traffic safety, 2–4
 - vision-based systems, 11

Intensity-based map, 366, 369, 389–394
Interaction between vehicles, 1285–1286
Interactive local hazard warning, 1238–1240
Internet, 1093, 1095
Intrinsic mode function (IMF), 961–964
Intrusion detection, 1175, 1209–1215
In-vehicle systems, 1169
Inverse depth parameterization, 372
IP address, 1099, 1144
IR vision systems, 40
ISA. *See* Intelligent speed adaptation system
ISO 15622, 615–617, 620, 625, 627
ISO 22179, 615, 618, 622
ITM algorithm. *See* Instantaneous topological map algorithm
ITS, 1553
IVHS. *See* Intelligent vehicle highway systems

J

J2735, 1228, 1236
J2945, 1228, 1238
Jamming, 1180–1181, 1189
Joint probabilistic data association filter (JPDAF), 1406

K

Kalman filter (KF), 373, 377, 382, 391, 403–404, 1313–1315, 1328
Kalman snakes, 1076
Key update, revocation, misbehaving, 1256–1257
KF. *See* Kalman filter

L

Landmarks, 367, 369, 371–373, 380, 389
Lane change assist (LCA), 211
Lane change decision aid system, 730, 733
Lane change warning, corner radius, 730, 733, 736, 742, 745, 753–755
Lane departure control, 519, 526, 528–530
Lane departure warning (LDW), 498, 514, 526, 528–530, 690–693, 696, 699, 704–706, 1000, 1023, 1558
Lane detection, 377, 1022–1030
Lane estimation, 373, 377, 378, 380
Lane keeping, 497, 505, 529, 1560, 1561
Lane-keeping aid (LKA) system, 211
Lane keeping assistance system (LKAS), 1063, 1064
Lane keeping support, 691
Lane position, 896, 898, 900, 901, 904–906, 908, 910
Lane-tracking, 377

Laser, 368–370, 377, 380, 381, 386–388, 394
Laser-scanners, 1037–1038
Lateral control, 1543
Lateral driving model, 149
Lateral dynamic, 242, 259–273
 – control systems, 271
Lateral force, 264–273
Lateral guidance, 730, 733, 753, 756
LCD, 480
LDW. *See* Lane departure warning
Lead acid battery, 293
Legal automation, 1528
Legal autonomous driving, 1522–1536
Level of intervention, 495, 512, 513, 517, 520, 531
Liabilities, 1522, 1530–1534, 1536
Light Detection And Ranging (LIDAR) sensors, 1064
Linear motion model, 1396
Link budget, 1100, 1102
Literature, 261
Lithium-ion battery (Li-Ion), 293
LKAS. *See* Lane keeping assistance system
Local area networks, 1099
Local dynamic map (LDM), 42, 43, 71
Local groupcast service, 1248–1249
Local high-speed unicast service, 1247–1249
Local sphere, 1094, 1095, 1120
Location-based map, 368, 369, 390, 394
Longitudinal control, 168–189, 642–648, 650, 1543–1546
Longitudinal dynamic by-wire control systems, 252–259
Longitudinal dynamic systems, 242–250
Longitudinal force, 263–265
Long-range radar (LRR) sensor, 37, 38
Long-term memory (LTM), 1363
Loosely coupled system, 459, 460
Low-speed unicast service, 1250–1251
Luminance, 1014

M

MAC. *See* Medium access control
Machine perception, 661–666, 669, 672, 678, 680, 683, 685
Machine screw actuator, 45
Magnetic nails, 700
Mahalanobis distance, 775
MAIS, 718, 723
Managing speeds of traffic on European roads (MASTERS), 595–597

- Mandatory, 584–589, 593–600
 - Mandatory lane changes (MLC), 550–551
 - Man-machine interface, 495, 520, 524, 525, 531
 - Manual on universal traffic control devices (MUTCD), 1125, 1137
 - Map, 464, 466–469, 471, 473–477, 479, 480, 482, 484–486
 - Map matching, 399, 453–455
 - techniques, 356, 360
 - Marginalized particle filter (MPF), 404
 - Markov chain Monte Carlo (MCMC) approach, 1408, 1409
 - Markov hypothesis, 1314
 - Markov random fields (MRF), 1428
 - MASTERS. *See* Managing speeds of traffic on European roads
 - Matching process, 1363
 - Matlab simulink-based program, 288
 - Maximum speed, 584, 585, 592
 - Mean speed, 586, 592
 - Medium access control (MAC), 1103
 - Mental workload, 522–523
 - Message insertion, 1181–1182
 - Metallic object detection, 38
 - Metric, 723
 - Metric models, 1318
 - Microscopic, 588
 - Midblock dash, 725
 - Millimeter wave (MMW) radars, 1391
 - MINIFAROS, 38
 - Minimum detectable velocity (MDV), 422
 - Misbehaving vehicle, 1195, 1196, 1201, 1205–1206, 1211
 - Mixed urban driving, 1386
 - MLC. *See* Mandatory lane changes
 - MMPF. *See* Multiple model particle filter
 - Mobile millennium, 1156
 - Mobile robot navigational safety
 - applications
 - off-road navigation, 1453, 1454
 - roadway and urban navigation, 1451–1453
 - constraints, 1436
 - hierarchical approach, 1436
 - iterative motion planning
 - ego-graphs, 1449–1450
 - input space sampling, 1444–1445
 - potential field techniques, 1443
 - RDT algorithm, 1450–1451
 - recombinant state lattices, 1451
 - state space sampling, 1446–1449
 - motion safety
 - conservative models, 1441
 - deterministic models, 1441
 - evasive trajectories, 1442–1443
 - ICS, 1437–1439
 - ICS approximation, 1442
 - limited decision time, 1439
 - lookahead, 1440
 - probabilistic models, 1441–1442
 - reactivity, 1442
 - τ -safety, 1442
 - space-time model, 1440
 - nomenclature, 1437
 - Mobility, 1540, 1546–1553
 - service, 1540, 1547
 - Model, 84–87, 89, 95, 98, 99, 102, 105, 107–111
 - Modeling of roads, 374
 - Model predictive control (MPC), 84–87, 109, 204–205
 - Monte-Carlo method, 725
 - Motion compensation, 1006
 - Motion models, 381
 - Motivators-rational for deployment of autonomous vehicles, 1288, 1290, 1305
 - Moving object detection, 760, 774
 - MPC. *See* Model predictive control
 - Multi-agent systems, 88, 89
 - Multilayer piezo actuators, 49
 - Multi-level control, 91
 - Multimodality, 1547
 - Multipath, 1100–1103, 1130, 1142, 1143
 - Multiple applications, 1252–1253, 1263, 1264
 - Multiple hypothesis tracking (MHT)
 - algorithm, 1405
 - Multiple model particle filter (MMPF), 429, 430
 - Multi-sensor fusion, 1040–1041
 - Multi-sensor systems, 1001
 - Multi-target tracking (MTT), 425
 - MUTCD. *See* Manual on universal traffic control devices
- ## N
- National Automotive Sampling System
 - Crashworthiness Data System (NASS CDS), 872–873
 - National Automotive Sampling System–General Estimates System (NASS/GES), 943

- National Highway Traffic Safety Administration (NHTSA), 944
 - National Institute of Health (NIH), 944
 - Naturalistic Driving Research, 560–577
 - Naturalistic driving study
 - data analysis, 569–570
 - data reduction, 566–569
 - Navidata, ADAS
 - advantages and disadvantages, 890, 891
 - camera and radar systems, 882
 - curve speed warning, 888–889
 - digital maps, automotive industry, 882–883
 - electronic horizon system, 890
 - energy-management application
 - eco-driving/predictive gear-shifting/predictive cruise control, 885–886
 - electric vehicle range prediction, 884–885
 - fuel-efficient routing, 883–884
 - GPS positioning, 890
 - overtaking assistant, 887–888
 - predictive front lighting, 886–887
 - reliable vehicle connectivity, 891
 - road geometries, 889–891
 - sensor fusion, 891
 - three-dimensional road model, 889
 - Navigation, road-bound vehicles
 - IMU, 417–421
 - odometric approach
 - accelerometers and gyros, 417
 - angular velocities, 415
 - GPS information, 417
 - lateral dynamics and curve radius, 415, 416
 - map-aided positioning, 415, 416
 - with parameter adaptation, 417, 418
 - slip-free motion, 415
 - wheel speed information, 416, 417
 - support sensors, 402
 - Navigation task, 608
 - k-Nearest neighbor classification method, 967
 - Nearest neighbor (NN) method, 1405
 - Near infrared (NIR) lamp, 1002, 1003
 - NEDC. *See* New European Driving Cycle
 - Nested logit model structure, 551
 - Network, 1128, 1143, 1144
 - centric vehicles, 1293, 1295
 - Neural networks, 89, 1054–1056
 - New Energy and Industrial Technology Development Organization (NEDO), 294
 - New European Driving Cycle (NEDC), 289, 302, 303
 - Nickel-metal-hydride battery (NiMH), 293
 - Night vision assistant, 799, 800, 807, 811
 - NIR lamp. *See* Near infrared lamp
 - NLOS conditions. *See* Non-line-of-sight conditions
 - Non-common mode errors, 440
 - Non-holonomic constraint, 452, 453
 - Nonlinear filter, 400
 - Non-line-of-sight (NLOS) conditions, 423–424
 - n*-point averaging, 964
 - Nuisance alarms, 978, 984
- O**
- OBD. *See* Onboard diagnostics port
 - OBD-II. *See* On-board diagnostics
 - Object and people detection
 - classification score, 1420
 - conditional probability, 1419
 - CRFs, 1418–1419
 - disadvantage, 1418
 - 2D scan data, 1418
 - precision-recall curves, 1420
 - Object detection, 625–630
 - Objective safety, 122
 - Object level tracking
 - clustering module, 1413
 - FCTA, 1413
 - grid-based clustering, 1413
 - lidar, 1414, 1415
 - re-clustering and track merging, 1413
 - track management, 1414
 - velocity estimation, 1413
 - Object uncertainty, 72, 73
 - Obstacle detection
 - algorithms design, 1036
 - brightness variations, 1035
 - features' complexity, 1035
 - flat ground hypothesis, 1034
 - methods
 - active sensor use, 1038
 - multi-sensor fusion, 1040–1041
 - vision, 1038–1039
 - moving behavior, 1035–1036
 - sensors
 - active and passive sensors, 1036–1038
 - parameters, 1036
 - proprioceptive and exteroceptive sensors, 1036
 - urban and off-road scene, 1034, 1035
 - OBU. *See* On-Board Unit
 - Occupancy grid map, 368, 369, 386–389, 393
 - OCSP. *See* Online Certificate Status Protocol

Odometer, 443, 444, 446, 449, 452
 Off-road particle filter, 429, 430
 On-Board Diagnostics (OBD-II), 1113
 Onboard diagnostics port (OBD), 1149–1150, 1159
 On-Board Unit (OBU), 1105, 1117, 1119
 Online Certificate Status Protocol (OCSP), 1256
 On-road particle filter, 429
 Operating points, 723
 Optical flow, 1074, 1076, 1077
 Optical warning, 669
 Optimal, 86, 109–111
 – control, 84–89, 101, 109–111, 113
 Optimal velocity model (OVM), 545
 Optimization, 710, 712, 715, 716, 723–725
 Overloading, 584, 589, 600
 Oversee, 1231, 1264
 Overtaking assistant, 887–888
 Owner's manual, 730, 753

P

Parameter learning, 1320
 Park Assist, 830–863
 Parking space measurement, 830, 850–851, 857
 Park steering control (PSC), 857, 858
 Park steering information, 845, 850, 851, 857
 Partial automation, 1527, 1532
 Partial (semi) control, 513, 514, 517, 531
 Partial differential equation, 1343
 Particle filter (PF), 404
 Passenger vehicles and autonomous vehicles, 1279,
 1285, 1287, 1292, 1297, 1305
 Passive pedestrian safety, 791–796, 820, 824
 Passive safety, 661, 666, 669, 710–716, 718, 720, 721,
 723, 726
 Passive sensors, 1037, 1063, 1064
 PATH, 86, 89–93, 95, 96, 112
 Path prediction, 632–633, 636
 Pattern analysis, 1074–1075, 1087
 Pedestrian
 – accidents, 788–791, 802, 805, 808, 823, 825, 826
 – classification, 760, 774–776
 – dummy, 779, 816
 – protection, 760, 773–780
 – protection system, 664
 – protection test facility, 790, 815–817
 – safety system, 659
 – zone driving, 1386
 PEMS. *See* Portable emission measurement system
 85th percentile speed, 592
 Perception functions, 1288, 1296

Perception-response, 493, 495, 496, 498–500, 503,
 505, 506, 531, 532
 PERCLOS, 899, 904–906, 909
 Performance and reliability, 609
 Personal area network, 1094
 Personal rapid transit (PRT), 1549–1551, 1556,
 1568–1574
 Personal sphere, 1094, 1120
 Perspective-based local descriptor, 1365
 PHD. *See* Probability hypothesis density
 Physiological, 494, 510, 516, 531
 PID, 84, 85
 – control, 84, 85
 Piezo bimorphs, 50
 Piezoelectric effect, 831–832
 Platoon, 83, 84, 89, 91–93, 95, 96, 101, 102, 109–113,
 170, 171, 173–182, 187, 1566, 1574
 PND. *See* Portable/Personal navigation device
 Point mass filters (PMF), 404
 Point of diminishing returns, 713, 715, 718
 Point of no return, 710, 716
 Points-of-interest (POI), 882
 Pollution, 1540–1541, 1550
 Portable emission measurement system (PEMS),
 285–287
 Portable/Personal navigation device (PND), 467,
 470–473, 480, 1148, 1150
 Position estimation, 322–323
 – GNSS receivers, 337
 – least-square (LS)
 – coordinates, 327
 – DOP, 328, 329
 – iterations, 328
 – linearization, 327, 328
 – pseudorange, 328
 – MMSE, 329
 Post-crash support systems
 – ACN system (*see* Automatic crash notification
 system)
 – information shared and dispatch
 protocols, 867
 – injury risk
 – crash attributes, 872
 – deltaV categories, 876–877
 – injured *vs.* uninjured US Tow-away crash
 populations, 876
 – ISS 15/ISS 16 injury targets, 877
 – lower injury severity threshold, 877
 – MAIS3+ injury, 873, 875
 – NASS CDS, 872–873

- URGENCY algorithm (*see* URGENCY algorithm)
- URGENCY parameters, crash direction, 873, 874
- in-vehicle system, 868
- moderate-to-high severity crash, 866
- opportunities
 - cellular telephones, 869
 - crash information, 878
 - emergency medical services, 869
 - enhanced data usage, 878
 - enhanced trauma criteria implementation, 878
 - fatality, 869
 - field triage decision scheme, 870, 871
 - high suspicion of injury, 870
 - injury model improvement, 878
 - paramedic judgment, 870, 872
 - trauma center care, 870
 - URGENCY instantaneous algorithm, 872
- PSAP, 867
- Telematics Control Unit, 868
- TSP, 866
- verbal communication, 869
- Power-split hybrid vehicles, 295
- Praxitele, 1566
- Preciosa, 1231
- Preconditioning, 718, 719, 722
- Prediction, 1462–1467, 1471, 1472, 1477
- Predictive cruise control, 885–886
- Predictive front lighting, 886–887
- Predictive gear-shifting, 885–886
- Prefill, 669, 671
- Preserve, 1231, 1240
- Preset, 669
- Pretensioning, 716
- PReVENT, 89, 94–96, 100, 112, 113
- Primary driving task, 608
- Privacy, 1175, 1178, 1179, 1182–1188, 1190–1194, 1197, 1207, 1215, 1219–1265
- Privacy against the CA, 1262–1263
- Privilege, 1224–1225, 1234, 1236, 1238–1240, 1242, 1246, 1247, 1249–1251
- Probabilistic model motion
 - collision risk estimation
 - behavior recognition and modeling, 1487–1489
 - chosen implementation, 1492–1494
 - conformal transformation, 1495, 1496
 - Gaussian process deformation model, 1490–1492
 - implementation issue, 1492
 - one vehicle trajectory, 1499, 1500
 - overall architecture, 1485, 1486
 - probability distribution, 1496, 1497
 - property, 1492
 - real-world coordinates, 1497, 1498
 - risk aggregation, 1500–1502
 - risk with driving behavior, 1502, 1503
 - collision warning, 1482–1483
 - driving simulation
 - adjacent line travel, 1512, 1513
 - confusion matrix, 1512–1514
 - 2D and 3D view, 1509
 - experimental setup, 1509, 1510
 - Gaussian Process, 1512
 - linear motion, 1511, 1512
 - mean and variance, 1514, 1515
 - red and yellow vehicles, 1510, 1511
 - risk values, 1514
 - simulator, 1509, 1510
 - ego-vehicle, 1482
 - intuition, 1484–1485
 - Monte Carlo simulation validation
 - experimental setup, 1504, 1505
 - HMM, 1506
 - mean and variance, 1506–1508
 - risk value, 1505, 1508
 - organization, 1485
- Probabilistic risk, 1466
- Probability hypothesis density (PHD), 390, 391
- Probe car, 465, 468, 478
- Probe data, 1115, 1117–1118
- Probe data collection, 1187, 1207, 1208, 1212
- Probes and intelligent vehicles
 - active signal management, 1158–1159
 - network design and maintenance, 1160–1161
 - privacy
 - encryption techniques, 1169
 - increasing probe penetration, 1170–1171
 - public perception, 1170
 - series of location points, 1170
 - smartphone tracking, 1170
 - probe maps
 - advantages, 1164
 - application development and validation, 1169
 - attributes, 1165

- behavioral maps vs. physical maps, 1164, 1165
- characteristics, 1164
- fuel economy application, 1166
- instantaneous speed data, 1167–1169
- intersection geometry, 1166–1167
- probe data maps drivers' behavior, 1163
- road classification, 1166
- virtuous cycle, 1165
- real-time applications and penetration requirements, 1153, 1154
- technology
 - aftermarket connections, vehicle buses, 1149–1150
 - aftermarket vehicle appliances, 1150
 - attributes list, 1149, 1150
 - communications systems, 1152–1153
 - data collection, 1149
 - high-end vehicles, 1149
 - OEM-installed systems, 1148, 1149
 - personal mobile devices, 1150–1151
- traffic control attribute detection
 - delay profile, 1161, 1162
 - green wave, 1161, 1163
 - light synchronization, 1163
 - metering lights, freeways, 1161, 1163
 - speed data, 1161, 1162
 - stop signs and stop lights, 1161
- traffic reporting
 - cell phone signals monitoring, 1156
 - communications costs, 1154
 - electronic toll collection systems, 1156
 - fixed infrastructure, 1154
 - GPS vs. cellular probes, 1156–1157
 - historic probe data, 1157
 - Inrix's fleet of probes, 1154, 1155
 - mobile millennium, 1156
 - real-time traffic information, 1153
- traffic smoothing, 1157–1158
- transportation probes
 - collection and communications systems, 1147
 - definition, 1147
 - PND, 1148
 - probe data, 1148
 - V2V applications, 1147
- weather, 1159

Product liability, 1526, 1530–1534

PRORETA project, 772

Proton exchange membrane (PEM) fuel cells, 298

Prototyping driving assistance systems, 573

Provider, 1220, 1221, 1226, 1243–1251, 1258, 1261

Provider service identifier (PSID), 1110, 1117

Proxemics, 1473, 1474

PRT. *See* Personal rapid transit

PSC. *See* Park steering control

Pseudo-random noise (PRN) codes, 319, 321

Pseudorange

- error sources
 - atmospheric errors, 335
 - ionosphere, 334
 - local path and multipath effects, 334
 - nonideal signal transmission, 333, 334
 - rough error budget, 334, 335
 - troposphere, 334
- measurement
 - carrier phase tracking, 331–333
 - code phase tracking, 331–332
 - navigation data, 333
 - signal acquisition, 330–331
- and position relation, 322–323
- relation and received signal
 - random distortion, 324, 325
 - systematic distortion, 325–327
 - undistorted signal, 324

PSID. *See* Provider service identifier

Public-key, 1221, 1222, 1224, 1228, 1231, 1233, 1234, 1237, 1238, 1240–1242, 1245, 1246, 1249, 1250, 1253–1255, 1259

Public key infrastructure, 1179

Public transit, 1568, 1569

Q

Quadratic program, 380

R

Radar, 366, 368–370, 380, 381, 384, 386, 388, 389, 393, 394, 1037

Radar sensor

- LRR, 37, 38
- SRR, 36–37

RAIM. *See* Receiver autonomous integrity monitoring

Range, 1093–1101, 1104–1110, 1120, 1124, 1125, 1130, 1131, 1133–1135, 1137–1139, 1141, 1143, 1144

Rao-Blackwellized particle filter (RBPF), 1406, 1408

Rapidly exploring dense trees (RDT) algorithm, 1450–1451

Rapid transit, 1556, 1568–1574

Rasmussen, J., 605

- Real world benefit, 807, 808, 820–826
 - Real-world in-car measurements
 - air conditioning and heating system, 286
 - emission components, 286
 - flow diagram, 287, 288
 - gas concentration measurement, 286
 - output data, 288
 - quality *versus* test bed measurements, 285
 - ultra-compact PEMS, 286, 287
 - Rear Vehicle Monitoring System, 742, 747–748, 754
 - Rear view camera, 845, 850, 852–855
 - Rearward-looking sensors, 670, 671
 - Rear wheel steering, 769–772
 - Received signal strength (RSS), 423
 - Received signal strength indication (RSSI), 1103
 - Receiver autonomous integrity monitoring (RAIM), 443
 - Red light running (RLR), 1128, 1129, 1141
 - Reflections and glares, 1008
 - Regional sphere, 1094, 1095, 1120
 - Regulatory law, 1522–1530
 - Re-keying, 1193, 1195, 1196, 1199, 1200, 1203, 1204, 1209, 1214, 1215
 - Remote sensing, 890
 - Replay relevance, 1236, 1238, 1240, 1242
 - RESPONSE Code of Practice, 610
 - Risk-based navigation, 1473
 - Risk compensation, 119
 - Road-bound vehicles
 - Bayesian posterior distribution, 400
 - dynamic map matching, 399, 400
 - Gaussian distribution
 - bimodal Gaussian distribution, 407, 408
 - covariance matrix, 406
 - four-way intersection, 406
 - Gaussian mixture distribution, 407
 - road-assisted navigation, 405
 - likelihood function, 400, 401
 - localization, 399
 - Manhattan problem, 407
 - manifold filtering, 401
 - map handling
 - computational issues, 413–414
 - shape format, 412–413
 - navigation (*see* Navigation, road-bound vehicles)
 - nonlinear filtering
 - definition, 402
 - finite state space models, 405
 - Kalman filter variants, 403–404
 - nonlinear motion model, 402
 - point mass and particle filter variants, 404
 - posterior density algorithms and representations, 403
 - posterior distributions, 405–406
 - state noise constraint, 401
 - tracking (*see* Tracking, road-bound vehicles)
 - two-dimensional motion models
 - complete Matlab algorithm, 409–410
 - dead-reckoning model, 408–409
 - manifold model, 411–412
 - tracking model, 411
 - virtual measurement, 401
 - Road map, 366, 367, 369, 373–386
 - Road prediction, 366–394
 - Roadside control, 111
 - Road traffic liability, 1530–1531
 - Robot coordinate frame, 1387
 - Robot freeway driving, 1386
 - RobuCab, 1371
 - Rolling horizon, 86, 87
 - Rotation matrix, 1493
 - Route guidance systems, 21–23
 - RSSI. *See* Received signal strength indication
 - Rule-based systems, 88
 - Rumble strips, 586
 - Run-off-road (ROR) crashes, 210
- ## S
- SAE. *See* Society of automotive engineers
 - SAE J2735, 1115, 1117
 - SafeCar Project, 597
 - SafeSpot, 89, 94–96, 112
 - SAFESPOT test bench
 - application integration, facility, 155, 157
 - application platform testing, 157
 - application unit testing, 155
 - cooperative vehicles development, 154
 - external vehicles and roadside units, 157
 - four interfaces, 155, 156
 - MARS, 148, 155
 - SAFEPROBE, 154, 155
 - SMA, 154
 - Safe state of a vehicle with the system(s), 278
 - Safety, 621, 647–650, 1540–1545, 1552
 - application, 1176, 1177, 1179–1181, 1186, 1187, 1207, 1215
 - comfort, 495, 497, 531, 606–611
 - critical event triggers, 566, 567, 569
 - distance, 1544, 1545

- Safety critical events, types, 568
- Safety Margin Assistant (SMA), 154
- Safety performance prediction, 724
- Satellite-based augmentation systems (SBAS), 338–340, 442
- Satellite positioning geometry
 - clock offset, 317, 318
 - clock synchronization, 317
 - geometrical range, 316, 317
 - intersection and uncertainty volume, 318, 319
 - pseudoranges, 317, 318
 - signal traveling time, 316, 317
- Saturation, 1017
- SBA. *See* Simulator brake actuation
- SBAS. *See* Satellite-based augmentation systems
- Scale invariant feature transform (SIFT)
 - descriptor, 1365
- Scaling series particle filter (SSPF), 1408
- Scanners, 211
- SCH. *See* Service channel
- Seamless ticketing, 1547
- Seat belt pretensioner, 669–671
- Secondary tasks, 608
- Security management services, 1220, 1228, 1253–1263
- Semantic sensor web (SSW), 68, 69
- Semi-automatic parking, 830, 846, 851, 857–859
- Sensing, 497, 498, 500, 504–506, 513–515, 521, 531
- Sensing systems
 - general in-vehicle sensors
 - accelerometer, 35
 - steering angle sensor, 35, 36
 - wheel speed sensor, 35
 - yaw rate sensor, 34
 - perception sensors
 - laser scanners, 38, 39
 - radar, 36–38
 - ultrasonic sensors, 40–41
 - vision systems, 39, 40
 - virtual sensors
 - digital map, 41–42
 - wireless communication, 42–44
- Sensor, 465, 466, 468, 472, 473, 478, 483, 485
- Sensor fusion, 380, 767, 780
- Sensor models or observation models
 - definition, 1392
 - dynamics model, 1396
 - physical model, 1393
 - pseudo-position, 1393
 - virtual sensors
 - black object detection, 1393, 1395
 - 3D to 2D scan conversion, 1393, 1394
 - ground filtering, 1393, 1395
 - image matching, 1395
 - image transform, 1395–1396
- Sensor systems, 787, 796–801, 808, 809, 820, 821, 826
- Sensotronic brake control (SBC), 53–54
- Sequence numbers, 1223
- Service channel (SCH), 1111, 1117, 1119
- Sevecom, 1231, 1232, 1258
- Shockwave damping
 - field experiments, 159–160
 - motion communication, 159
 - motorway network, Netherlands, 159, 161
 - prevention/mitigation, 158–159
 - speed diagrams, equipment rates, 160, 161
 - traffic flow characteristics, 158
- Short-lived certificate, 1205
- Short-range radar (SRR), 36–37
- Short-term memory (STM), 1363
- Shutter, 1017
- Side Assist, 742, 752–754
- Side Blind Zone Alert, 746–747, 754
- Side Collision Prevention (SCP), 750, 754
- Sight distance, 1125, 1129, 1130, 1134, 1140
- Sigma-point filters (unscented Kalman filters), particle filters, 458
- Sigmoid function, 762, 777
- Signal acquisition, 330–331
- Signal phase and timing (SPAT), 1119, 1127
- Signal-to-noise, 1100, 1102, 1141
- SimTD, 1228, 1231
- Simulation approach
 - application, 142
 - application testing and verification
 - laboratory test environment, 153, 154
 - SAFESPOT, 154–157
 - WiFi communication, 153
 - application validation and evaluation (*see* Shockwave damping)
 - concept development
 - CACC, 150–151
 - definition, 143
 - ITS function and application, 150
 - product design, 150
 - results, 151–152
 - setup, 151
 - decision process, 142
 - development process, 142–144

- function, 142
- requirements, 144, 145
- simulation environment
 - communication, 146
 - high level of control, 147
 - post-processing output, 146
 - R&D tools and tests, 146
 - traffic, 145, 146
 - vehicle and application, 145
- system, 142
- system complexity, 141
- traffic flow
 - acceleration, 148, 149
 - conducting process, 147, 148
 - driver behavior, 149
 - driver model, 147, 148
 - mathematical specification, 148
 - microscopic concept, 147
 - random process, 147
 - statistical analysis, 147–148
- Simulator brake actuation (SBA), 256–258
- Single sided braking, 769–771
- Situational awareness (SA), 366–394, 589
 - advanced driver support functionality, 63–64
 - and communication
 - failure modes, 68
 - high-speed wireless, 67–68
 - interoperability, 68
 - onboard redundancy, 68
 - ontology, 69
 - SSW, 68, 69
 - ubiquitous, 68
 - conceptual system decomposition, 64
 - control
 - architecture, 75
 - data interpretation algorithms, 78
 - layers, 75, 76
 - longitudinal control, 75–77
 - primary signal and secondary information flows, 76
 - speed advice application, 74–75
 - control functionality, 64
 - decision making, 65
 - defense and mobility domain, 67
 - definition, 65
 - functional levels, 62
 - JDL model, 67
 - model, dynamic decision making, 65, 66
 - primary functions, 63
 - situation assessment, 65, 67
 - state of knowledge, 65
 - *World Model*, 64–65
 - world modeling, representation
 - advantages, 70
 - conceptual objects, 71
 - configuration, 70
 - geometry, 71
 - hierarchy, inheritance, 71
 - LDM, 71, 72, 74
 - object relations, 73
 - position, motion state, 71
 - SAFESPOT Integrated Project, 70–71
 - signal flow process, 69
 - uncertainty, 73, 74
- Situation analysis, 760–765, 767, 768, 773, 776, 777, 780
- Situation assessment, 665
- S. Krauss (S-K) model, 547
- Slow eye closures, 899, 900, 904, 908
- Smartbus–2000, 1569
- SmartTer platform, 1452
- Smartway project, 1569
- Smear effect, 1006–1008
- Sneakernet, 1149, 1152
- Sobel phase analysis
 - edges detection, 1051
 - phase distribution, 1051–1053
 - sign translation and rotation, 1051
 - Sobel edges and Hough images, 1050
 - supervised learning methods, 1050
- Society of Automotive Engineers (SAE) J2735, 1132, 1135, 1228, 1236, 1238, 1240
- Space segment, 320, 321
- Spacing policy, 99, 100, 102, 107, 112
- SPAT. *See* Signal phase and timing
- Spatial road map, 344–346
- Spectrum of assessment, 722
- Speed, 120, 121, 123–125, 127–129, 133, 134
- Speeding, 583, 584, 586–590, 594–600
- Speed management, 590, 600
- Stabilization, 608
- Standard of quality, 723
- STARDUST, 594–595
- Start-inhibit system, 1001
- State estimator, 457
- State space sampling
 - boundary state sampling techniques, 1448–1449
 - initial state variation, 1446, 1447
 - inverse trajectory generation problem, 1446

- model-predictive trajectory generation, 1447, 1448
- motion control, 1446
- search space generation, 1446
- trade-off, 1446
- Static event, 1125–1126
- Static local hazard warning, 1237–1238
- Steer angle actuator, 769, 770
- Steer by wire feedback design, 243, 260–262
- Steering actuator, 763
- Steering distance, 761–763
- Steering encoder, 444, 452
- Steering intervention, 760, 770, 771, 780
- Steering torques, 756
- Steering wheel
 - Ackermann steering angle, 213–214
 - critical speed, 215
 - inputs, 900–901
 - rack-and-pinion linkage, 213
 - steering gearbox, 213
 - tire slips, 214
 - understeer gradient, 214–215
- Steer torque actuator, 768–771
- Stereo vision, 760, 768, 774
- Stereovision systems, 40
- Stochastic mapping algorithms, 368
- Stochastic variables, 724
- Stop & go, 1560
- Stopping criterion, 151–152
- String
 - stability, 87, 93, 96, 98, 100–106, 108, 112, 172, 173, 175, 177, 178, 184
 - of vehicles, 96–98, 102
- Structure-from-Motion (SfM), 1358
- Structure learning, 1320, 1326
- Subjective safety, 121, 130, 131
- Subsampling, 1016
- Symmetric, 1221, 1224, 1231, 1239, 1240, 1247–1250, 1252–1256
- Synchronization, 1001
- System architecture, 672–676, 683
- System engineering V-model, 143
- Systems engineering process, 571
- System status diagram, 733

T

- Tactical stage lane changing models
 - Ahmed model
 - forced merging behavior, 552–553
 - gap acceptance process, 551–552

- lane changing decision process, 551, 552
- MLC and DLC, 550–551
- Gipps' model, 550
- Hidas model, 554–556
- MOBIL model, 556–557
- Wiedemann and Reiter model, 553–554
- Tampering, 1190, 1209, 1211, 1215
- Target braking, 671
- Target selection, 626, 628, 630–637, 642
- Target speed, 590, 591, 596
- TCP, 1144
- TCP/IP, 1104
- Technological classification of intelligent vehicles, 1286–1288
- Telematics service provider (TSP), 866
- Template matching, 1075, 1080, 1081
- Testing and evaluation, 495, 525–526, 532
- Testing driver assistance systems, 573
- Test procedure, 736
- Thermal shadows, 1006
- Third-party provability, 1221
- Threat agent, 1175, 1177–1183, 1188, 1189, 1209, 1215
- Threat model, 1175–1215
- Threat motive, 1178
- Three-level model hierarchy (Bernotat), 608
- Throttle, 585, 586, 595, 600
- Tightly coupled system, 460
- Time constraints/geographic constraints, 1224
- Time gap, 616, 617, 620, 622–629, 632, 638–642, 651
- Time headway, 100, 102, 103, 107–109
- Timeline of autonomous ground vehicles, 1298
- Timestamping, 1223
- Time-to-brake (TTB), 765, 775, 777
- Time-to-collision (TTC), 1482, 1483, 1485
- Time-to-kickdown (TTK), 765
- Time-to-steer (TTS), 765, 776
- Time-to-x, 760, 777
- Tip to tip headway, 1545
- Tire models
 - brush model, 218
 - linear tire model, 219–220
 - magic formula tire model, 217–218
 - swift tire model, 219
- TLS. *See* Transport layer security
- Tolling, 1124, 1128, 1130, 1176, 1181, 1187
- Tones and voice messages, 981
- Top-down approach, 610
- Topological information, 455
- Torque vectoring, 769, 773

Tracking, 1182–1188, 1211, 1214

Tracking, road-bound vehicles

- radar system
 - electromagnetic waves, 421
 - false detections, 422
 - GMTI, 422, 423
 - GNN association algorithm, 422
- sensor network
 - acoustic power, 425
 - binary proximity sensor, 425
 - EPD filter, 427
 - microphone network, 427, 428
 - MTT, 425
 - road segment, microphones, and coordinates, 426, 427
- support sensors, 402
- vision sensors
 - azimuth and inclination angles, 429
 - Cartesian coordinates, 429
 - definition, 428
 - fail-safe algorithm, 431
 - global Cartesian reference system, 428
 - MMPE, 429, 430
 - negative information, 431
 - off-road PF, 429, 430
 - path of the car, 429
 - RMSE results, 429, 430
 - simulation environment, 429
 - target-tracking filter, 431
- wireless radio network
 - base stations and measurement locations, 424, 425
 - cumulative distribution functions, 424, 426
 - fingerprinting, 424
 - multipath and NLOS conditions, 423–424
 - Okumura–Hata model, 423
 - range measurements, 422
 - RSS measurement, 423–424
 - static and dynamic localization, 424–426

Traffic calming, 583

Traffic control, 83, 88, 93, 94, 96

Traffic jam, 588

Traffic management, 83, 113

Traffic safety, 583, 600, 606, 609

Traffic signal, 1136

Traffic sign recognition

- algorithm flowchart, 1045
- classification
 - bilinear interpolation, 1053
 - 256 bin gray scale histogram, 1053

- contrast stretching and filtering, 1054, 1055

- intensity histograms, 1053

- neural network, 1054–1056

- reduced and normalized regions, 1054

- tracking, 1056

- color analysis

- chromatic equalization, 1047–1048

- color segmentation, 1045–1047

- gray-scale and color cameras, 1044

- missed signs, 1044

- output and results

- empirical tests, 1057

- false positives, 1058–1059

- illumination conditions, 1056–1057

- road and junction structure, 1057

- rotated signs, 1058

- triangular and rectangular signs, 1057, 1058

- shape detection

- bounding boxes merge and split, 1049–1050

- pattern matching, 1050

- sobel phase analysis, 1050–1053

- sorting, 1049

- speed limit sign recognition, 1044

Trajectory generation, 765, 776, 777

Transport Layer Security (TLS), 1181, 1190

Transverse bars, 596

Trauma centers, 713

Travel time, 586–588, 595

Trilateration, 834–836, 847

Trust, 121, 126–128, 131

TTB. *See* Time-to-brake

TTK. *See* Time-to-kickdown

Two ray model, 1102

Typical behavior, 1463

U

UDP, 1104, 1142–1144

Ultrasonic parking aid, 831, 846–847, 849

Ultrasonic sensor, 831, 834–838, 846, 847

Ultrasonic transducer, 832, 833, 835, 836

Underloading, 589

Unlinkability, 1182, 1183, 1187, 1203

Unlinked certificate, 1192, 1197–1207

Untreated sleep apnea syndrome (SAS), 946

Urban driving cycle (UDC), 289

URGENCY algorithm

- capture rate, 874, 876

- crash populations, 873

- instantaneous algorithm, 872

- logistic regression models, 872

- User, 1221, 1225, 1226, 1232–1235, 1237, 1243–1246, 1251, 1253, 1261, 1262, 1264
 - centered design, 571
 - registration, 1253
 - segment, 320, 321

V

- Vacuum-operated throttle actuator, 51, 52
- Validation plan, 143–144
- VDTLS. *See* Vehicular datagram transport layer security
- Vectus, 1571
- Vehicle, 83, 89–109, 112
 - centric, 1275, 1286, 1288–1290, 1293, 1295–1300
 - coordinate system, 447–451
 - data collection, 565–566
 - driver control loop, 237, 239–240
 - dynamics, 437, 458–460
 - following, 95–99, 102
 - following control, 638–640
 - instrumentation, 565
 - model, 98, 99, 102, 107
 - motion sensors, road maps, 437
 - safety, 606, 610
- Vehicle-in-the-Loop (ViL), 677, 684, 685, 790, 811, 815, 817–819, 826
- Vehicle lateral and steering control
 - active safety systems, 211
 - components
 - steering wheel, 213–215
 - suspension system, 215
 - tires, 212–213
 - lane change maneuver, 211
 - lane-keeping support systems, 210, 211
 - ROR crashes, 210
 - vehicle model
 - bicycle model, 215–216
 - error coordinates, 220–221
 - global position, 221, 222
 - kinematic relations, 217
 - momentum, 216
 - Newton's second law, 216
 - slip angles, 217, 221
 - state space variables, 221–223
 - tire models (*see* Tire models)
 - yaw rate, 216, 221
 - vehicular safety enhancements, 210
- Vehicle safety, 4
- Vehicle safety communications applications (VSC(A)), 1230

- Vehicle safety communications consortium (VSCC), 1230
- Vehicle state variables
 - classification, 949–950
 - detection methods
 - advantages and disadvantages, 967–968
 - pattern classification schema, 955, 956
 - predefined mathematical model, 955
 - lateral position, 954
 - sensing, 949
 - vehicle speed, 954–955
 - vehicle steering activity
 - amplitude duration squared theta, 953
 - dozing off intervals, 951, 952
 - “impaired” phase, 951
 - jerky motion, 953
 - macro-corrections, 950
 - micro-correction, 950
 - steering correction, 951, 953
 - steering velocity, 953
 - steering wheel frequency, 951
 - steering wheel reversal rate, 951
 - weight flat zero, 954
 - yaw/brake/acceleration, 955
- Vehicle-to-Infrastructure (V2I), 1123–1144
- Vehicle-to-Vehicle (V2V), 1123–1144, 1147
- Vehicular datagram transport layer security (VDTLS), 1207, 1208
- Velocity encoders, 444, 446, 449
- Verification plan, 144
- Vertical force, 263
- VICS, 466–468, 477
- Video camera, 843–844, 861
- Vienna Convention on Road Traffic, 1523, 1561, 1562
- Vigilance, 523–525
- VIIC proof of concept, 1230, 1258–1260
- VII proof of concept, 1100, 1142
- Virus, 1181, 1189
- Vision and IMU data fusion
 - accelerometer and sensor in line, 1338
 - algorithm
 - camera frame, 1350
 - matrix computation, 1350, 1351
 - a priori known, 1350
 - rank matrix, 1351
 - speed determination, 1349
 - closed-form solutions
 - sensor measurements, 1346
 - without bias, 1347–1349

- considered system
 - with bias, 1341
 - camera point feature, 1339
 - 3D vector, 1339, 1340
 - multiple features, 1340–1341
 - observation function, 1340
 - quaternion, 1339, 1340
 - sensor assembling, 1339
 - 3D-SLAM, 1337
 - EKF, 1337
 - observability property
 - with bias, 1344–1345
 - with gravity, 1343
 - multiple features, 1344
 - necessary conditions, 1346
 - unknown gravity, 1345–1346
 - without gravity, 1342–1343
 - observable modes, 1338
 - performance evaluation, 1352–1353
 - rotation matrix, angular speed integration, 1353
- Vision-based ACC
- ADAS applications, 1062, 1064
 - ad-hoc vision algorithm, 1066
 - comfort deceleration, 1063
 - data fusion, 1064
 - field of view, 1066
 - flow chart, 1064, 1065
 - LIDAR sensor, 1064, 1065
 - LKAS, 1063, 1064
 - medium and far region, 1063
 - overtaken region, 1063
 - pre-crash warning system, 1064
 - principal categories, 1062
 - vehicle detection
 - clustering algorithm, 1066
 - edges, 1067
 - knowledge-based methods, 1066
 - LIDAR-based method, 1068
 - lights, 1067–1068
 - symmetry, 1067
- Vision-based topological navigation
- 3D reconstruction, 1356
 - environment representation
 - images, 3D and visual features, 1359, 1360
 - key images selection, 1362
 - visual memory, 1359
 - visual memory update, 1362, 1363
 - visual memory vertices, 1360, 1361
 - visual paths, 1359–1360
 - visual route, 1361–1362
 - weighted directed graphs, 1361
 - experimental setup, 1371
 - image retrieval, 1358
 - large displacement
 - autonomous navigation, 1374, 1377, 1378
 - evaluation, RTKGPS, 1378–1379
 - initial localization, 1374, 1375
 - learning step, 1373, 1374, 1376
 - loop closure, 1372, 1373
 - memory localization
 - global descriptors, 1364
 - hybrid descriptors, 1366
 - image acquisition, 1363–1364
 - local descriptors, 1364–1365
 - matching process, 1364
 - mobile robot environment, 1357
 - robot trajectory, 1356
 - route following
 - camera displacement, 1366
 - control design, 1370
 - control objective, 1367
 - trajectory-following task, 1367
 - vehicle modeling, 1367–1369
 - vision-based control scheme, 1367, 1368
 - visual servoing, 1366
 - SfM problem, 1358
 - SLAM problem, 1358
 - view-sequenced route reference, 1357
 - visual memory acquisition, 1357
- Vision sensor, 863
- Vision Zero, 710
- Visual, audio, and haptic, 504
- Visual display, 596
- Visual odometry, 369
- Voluntary, 585–589, 596, 598–600
- VSC3, 1230
- VSC(A). *See* Vehicle safety communications applications
- VSCC. *See* Vehicle safety communications consortium
- ## W
- Warning, 585, 586, 588, 589, 592, 597, 598, 600
- application, 1127, 1133–1135, 1137, 1138
 - closing vehicle warning, 730, 733, 734, 736, 739
 - compliance, 980–981
 - dilemma, 667, 668

- modalities, 584
- symbols, 979–980

Warping the suspension, 769

WAVE. *See* Wireless access in vehicular environments

Waveform design, 368

WAVE service advertisement (WSA), 1110, 1111, 1117, 1119

White balance, 1017

Wiedemann model

- operational stage acceleration models, 548
- tactical stage lane changing models, 553–554

Wireless access in vehicular environments (WAVE), 1108, 1110, 1111, 1142

Wireless communication, 198

- CALM, 44
- C2C-CC, 44
- DSRC, 42, 44

- ETSI, 44
- ITS applications, 42, 43
- types, 42

World frame, 1387

World Health Organization (WHO), 3

World modeling, 1281–1284, 1291, 1292, 1299, 1300

World Wide Web, 1095

WSA. *See* WAVE service advertisement

Y

Yaw rate sensor, 34

Z

Zero Emission Program (ZEV), 294

Z-transform, 102, 105, 106